

MÉMOIRE DE RECHERCHE

Département des Sciences du Langage

M2 Linguistique, Informatique et Technologies du Langage (LITL)

Transdisciplinarité et sous-spécification des têtes de segments des titres de documents scientifiques

Damien GOUTEUX

Sous la direction de Mme Josette Rebeyrolle et M. Ludovic Tanguy

2018 – 2019

[Introduction 5](#_Toc16260628)

[I. Exploration du corpus à la lumière de l’état de l’art 8](#_Toc16260629)

[I.1 Origine des données et prétraitement des données 8](#_Toc16260630)

[I.1.1 Récupération des données 8](#_Toc16260631)

[I.1.2 Étiquetage et analyse syntaxique en dépendances 9](#_Toc16260632)

[I.1.3 Segmentation des titres 10](#_Toc16260633)

[I.1.4 Sélection de la racine des segments 11](#_Toc16260634)

[A. Titres avec un segment et une racine 11](#_Toc16260635)

[B. Titres avec un segment et deux racines 12](#_Toc16260636)

[C. Titres avec un segment ayant une racine suivie d’un segment sans racine 12](#_Toc16260637)

[D. Titres avec un segment sans racine suivi d’un segment avec racine 12](#_Toc16260638)

[E. Titres avec un segment avec racine suivi d’un segment avec racine 12](#_Toc16260639)

[F. Algorithme de sélection de tête de segment 12](#_Toc16260640)

[I.2 Description des données et mesures du corpus 13](#_Toc16260641)

[I.2.1 Description des données des titres 13](#_Toc16260642)

[I.2.2 Sélection des données selon la structure et donc la nature des titres 14](#_Toc16260643)

[A. Structures des titres 14](#_Toc16260644)

[A.1 Titres composés d’un seul segment 14](#_Toc16260645)

[A.2 Titres composés de deux segments 14](#_Toc16260646)

[B. Nature des titres 15](#_Toc16260647)

[I.2.3 Mesures effectuées sur notre corpus de travail 18](#_Toc16260648)

[II. Caractérisation des têtes de segments 21](#_Toc16260649)

[II.1 Têtes de segments 21](#_Toc16260650)

[II.1.1 Les têtes de segments spécifiques 22](#_Toc16260651)

[A. Définitions théorique et opératoire 23](#_Toc16260652)

[Corrections de Talismane 24](#_Toc16260653)

[B. Résultats et évaluations des résultats 25](#_Toc16260654)

[II.1.2 Les têtes de segments transdisciplinaires 29](#_Toc16260655)

[A. Définitions théorique et opératoire 29](#_Toc16260656)

[B. Résultats et évaluations du résultat 29](#_Toc16260657)

[C. Remarques sur les sous-corpus 30](#_Toc16260658)

[Conclusion de la recherche de têtes spécifiques et transdisciplinaires 31](#_Toc16260659)

[II.2 Les noms généraux sous-spécifiés 32](#_Toc16260660)

[II.2.1 Principes théoriques 32](#_Toc16260661)

[II.3 Présence des constructions spécificationnelles dans notre corpus 35](#_Toc16260662)

[II.4 Caractéristiques retenues de rapprochement 36](#_Toc16260663)

[II.5 Points communs et différences avec les noms sous-spécifiés 37](#_Toc16260664)

[II.5.1 Facteurs de rapprochement 37](#_Toc16260665)

[II.5.2 Facteurs de différenciation 37](#_Toc16260666)

[II.5.3 Règle de rapprochement 38](#_Toc16260667)

[III. Schémas récurrents et constructions spécificationnelles 38](#_Toc16260668)

[III.1 Émergences de schémas fréquents récurrents 38](#_Toc16260669)

[III.1.1 Schémas intrasegmentaires 38](#_Toc16260670)

[III.1.2 Schémas sur deux segments 38](#_Toc16260671)

[III.2 Transdisciplinarité des schémas et schémas non centrés sur têtes 39](#_Toc16260672)

[III.2.1 Transdisciplinarité des schémas 39](#_Toc16260673)

[III.2.2 Schémas non centrés sur une tête 39](#_Toc16260674)

[III.3 Des schémas de constructions spécificationnelles ? 39](#_Toc16260675)

[IV. Discussion sur nos résultats, limites et perspectives 39](#_Toc16260676)

[IV.1 Limites de notre travail 39](#_Toc16260677)

[IV.2 Perspectives 40](#_Toc16260678)

[Conclusion 40](#_Toc16260679)

[A1. Références bibliographiques 40](#_Toc16260680)

[A2. Liste des têtes 44](#_Toc16260681)

[A.2.1 Liste des têtes spécifiques aux domaines 44](#_Toc16260682)

[A.2.2 Liste des têtes transdisciplinaires 65](#_Toc16260683)

[A3. Étiquettes utilisées par Talismane et HAL 66](#_Toc16260684)

[A3.1 Catégories morphosyntaxiques de Talismane 66](#_Toc16260685)

[A3.2 Relations de dépendance de Talismane 68](#_Toc16260686)

[A4. Éléments techniques 69](#_Toc16260687)

[A4.A Présentation de l’API de requêtage de notre corpus 69](#_Toc16260688)

[A4.B Analyse de 100 titres traités par Talismane 70](#_Toc16260689)

[A5. Index des tableaux 75](#_Toc16260690)

[A6. Index des graphiques 75](#_Toc16260691)

[A7. Index des notions mentionnées 76](#_Toc16260692)

# Introduction

Un titre de document scientifique est un énoncé singulier d’une importance cruciale. D’une part, il s’agit d’un texte très court d’une dizaine de mots. D’autre part, il constitue le premier contact entre le document et les lecteurs et, dans 92 % des cas, le lecteur en restera là (Mabe et Amin, 2002). C’est sur la lecture du titre seul, indépendamment du document titré, que le chercheur fait son tri parmi la littérature scientifique (Goodman et al., 2001). Sa production augmente constamment en doublant tous les 12 ans (Stix, cité dans Salager-Meyer et al. 2013). Ce tri par la lecture du titre soulève la question de l’information qu’il contient. On peut s’interroger sur les mots et les structures utilisés pour convoyer cette information. Cet intérêt s’est traduit par de nombreux articles sur les titres en anglais, mais les titres français ont été moins étudiés, on peut citer les travaux de Ho-Dac et al. (2001), Rebeyrolle et al. (2009) et Tanguy et Rebeyrolle (à paraître). Nous prenons en compte le titre uniquement dans sa fonction informationnelle, considérant qu’elle est la plus importante, soutenu en cela par Haggan (2004) et Hartley (2005). Cette dimension est également la plus facile à analyser. Nous laissons donc de côté la fonction d’attraction qui peut considérablement obscurcir son sens (Hartley, 2005) ou faire appel à des notions complexes pour le traitement automatique des langues comme l’humour (Sagi et Yechiam, 2008 ; Subotic et Mukherjee, 2014).

Nous devons en premier lieu revenir sur notre travail effectué durant la première année de master sur les titres de publications scientifiques en français. Nous avions étudié trois schémas fréquents dans un corpus de titres de publications scientifiques. Par schéma, nous entendons une séquence d’éléments pouvant comporter des choix entre plusieurs éléments et des répétitions d’éléments. Un élément peut être une classe grammaticale (nom - N, adjectif qualificatif - ADJ, préposition - P, conjonction de coordination CC…), un sous-ensemble d’une classe (nom commun - NC), un lemme (“et”) ou un signe de ponctuation (“;”). Par exemple, le schéma **NC ADJ** encode la séquence “un nom commun suivi d’un adjectif qualificatif”.

On dit qu’une séquence de mots et de signes de ponctuation dans un énoncé correspond à un schéma lorsqu’elle obéit à la séquence décrite. Les éléments décrits par le schéma peuvent être alors individuellement associés à un mot ou un signe de ponctuation de la séquence textuelle correspondante, on dit que les mots et signes peuplent le schéma. Ainsi la séquence de mots “Villes durables et changement climatique” correspond, entre autres, au schéma **NC ADJ CC NC ADJ** ainsi qu’au schéma **NC ADJ et NC ADJ**. Le premier n’utilise que des classes grammaticales comme éléments, le second utilise quatre classes grammaticales et un lemme, “et”, comme éléments. La séquence de mots “union parfaite ou mariage impossible” correspond au premier schéma mais pas au second. Pour les deux schémas, leur premier élément, NC, est associé au mot “Villes” pour le premier exemple et “union” pour le second.

Les trois schémas étudiés dans notre travail précédent étaient :

* un double point suivi d’un un syntagme nominal dont le nom est complémenté par un syntagme prépositionnel (: NC P NC),
* un double point suivi d’un syntagme prépositionnel dont le nom est complémenté par un syntagme prépositionnel (: P NC P NC),
* un double point suivi d’un syntagme nominal constitué de deux noms coordonnés (: NC CC NC).

Nous laissions la possibilité d’avoir des adjectifs qualificatifs pour les noms de chaque schéma mais par souci de simplification nous écartons cette possibilité ici. Dans notre corpus de 85 500 titres, le premier schéma couvrait 50% des titres, le deuxième 5% et le dernier 10%, soit une couverture totale de 65 % de notre corpus. Nous avions ensuite étudié les noms et les couples de noms les plus fréquents peuplant ces schémas. Nous avions constaté l’utilisation récurrente et transdisciplinaire de noms issus du vocabulaire du domaine scientifique, des noms abstraits, dont les 11 plus fréquents étaient :

* étude, cas, approche, analyse, application, pratique, exemple, enjeu, perspective, modélisation, limite.

Nous avions remarqué que ces noms sont des noms généraux tels que définis par Halliday et Hasan (1976), *“a small set of nouns having generalized reference”*, servant à maintenir la cohérence du texte. Lexicalement, ces noms appartiennent aux listes de noms généraux fréquemment employés dans un emploi sous-spécifié (NSS) telles qu’elles ont été définies par Schmid (2000) et Flowerdew et Forest (2015). Un NSS est un *« nom abstrait dont le sens complet peut seulement être spécifié en référence à son contexte* *»* (Flowerdew, 2006). Un point important est que le NSS est un emploi et non une nature lexicale, même si certains ont une appétence pour cet emploi (Schmid, 2000). Un NSS possède la particularité d’avoir un faible contenu sémantique et une très large application référentielle. La fréquence et la transdisciplinarité, qui plaident pour un faible contenu sémantique des noms que nous avions repérés, jouent en faveur d’un rapprochement avec les noms sous-spécifiés. Pourtant, l’utilisation de cet emploi, dont le trait caractéristique est un faible contenu sémantique du nom, soulève des questions dans un espace comme le titre où chaque mot est compté. De plus, l’emploi de noms de façon sous-spécifiée repose sur leur inclusion dans des constructions spécificationnelles (CS) (Legallois, 2008) qui mettent en rapport le nom sous-spécifié avec un contenu spécificationnel. Les deux CS les plus fréquemment étudiées (Legallois, 2008 ; Schmid, 2000) sont :

* **NSS** + [verbe être] + *proposition subordonnée complétive attribut du sujet* : “le **problème** est *que l’homme souhaite toujours plus*”,
* **NSS** + [verbe être] + **de** + *proposition subordonnée infinitive* : “le **problème** est **de** *délimiter nos souhaits*”.

Cette définition opérationnelle s’accompagne d’une définition fonctionnelle en trois points présentés par Schmid (2000) :

* Fonction sémantique : mise en perspective de tranches d’information variables.
* Fonction cognitive : formation de concepts temporaires dans des concepts nominaux uniques.
* Fonction textuelle : structuration du texte à travers la représentation de segmentation de nature variable.

La dernière fonction, qui rapproche les NSS du fonctionnement des pronoms, est difficilement perceptible dans un énoncé aussi court qu’un titre. Néanmoins les deux premières peuvent s’appliquer : que l’on qualifie de *problème* ou de *question* un élément du titre est une mise en perspective que le locuteur impose à son lecteur. De même, est regroupé sous ce NSS tout ce qui vient après dans le titre, formant un concept temporaire. Néanmoins, on ne retrouve pas les CS définies plus haut dans les titres.

La classe de nom ayant émergé de notre premier travail se rapprocherait-elle de ces NSS ? Pour cela, nous voulons étudier un ensemble de caractéristiques qui permettraient de les rapprocher des NSS. Et en l’absence de constructions spécificationnelles dans les titres, nos noms s’intègrent-ils néanmoins dans des énoncés correspondant à des schémas d’utilisation très fréquents ? Pour répondre à ces questions, nous utiliserons une approche se basant sur le traitement automatique des langues et la linguistique de corpus (Cori et David, 2008).

Nous pensons que la classe de nom ayant émergé dans notre premier travail peut gagner à être redéfinie par une autre approche, indépendante de sa position immédiatement après le double point. Nous avions écarté également dans notre précédente étude toute la partie avant le double point et les phénomènes récurrents pouvant y survenir, perdant ainsi des découvertes potentielles. Or, nous faisons l’hypothèse, soutenue par notre intuition et notre connaissance du précédent corpus, que le premier nom que nous étudions immédiatement après le double point est le noyau (ou tête) du syntagme de premier niveau du segment après le double point et donc la tête du segment. Nous redéfinissons donc notre cible comme les têtes de segment et nous élargissons cette étude, en ne regardant plus seulement le segment après le double point, mais également le segment avant. Nous élargissons également notre étude au titre à un seul segment et aux titres à deux segments séparés par un autre signe de ponctuation que le double point. Notre étude portera donc sur toutes les têtes, ou noyaux, des segments des titres. Dans l’exemple ci-dessous, le titre est constitué de deux segments, délimité par le double point, avec en gras leur tête :

(1) Titre n°1258625 : Un nouvel **OVNI** dans le ciel réunionnais : la **transparence** des prix

Tout d’abord, pour ce nouveau travail, il nous faut découper nos titres en segments en reprenant et en amendant une liste de signes de ponctuation qui segmentent les titres en anglais établie par Anthony (2001). Ensuite, pour trouver les têtes de syntagmes, plutôt que de simplement parcourir le segment et prendre le premier nom rencontré comme nous le faisions en première année, nous avons décidé d’utiliser l’analyse syntaxique en dépendances (Tesnière, dans Schwischay, 2001) qui produit un arbre dont la racine est la tête du segment. Une racine dans le cadre de l’analyse syntaxique en dépendances est un mot uniquement régisseur et jamais régi. Ce sont ces racines dont nous voulons étudier le rapprochement possible avec les noms sous-spécifiés. Nous voulons caractériser ces racines et les schémas récurrents dans lesquels elles s’insèrent, dans un corpus de titres de publications scientifiques.

Nous gardons à l’esprit l’existence de spécificités disciplinaires dans l’écriture des titres pour l’anglais (Haggan, 2004 ; Lewison et Hartley, 2005 ; Soler, 2007, 2011 ; Nagano, 2015), Tanguy et Rebeyrolle, à paraître, pour le français). Nous ne manquerons pas de déterminer dans le cadre de notre problématique s’il existe des variations des racines et des schémas suivant les disciplines.

Notre étude se déroulera en quatre temps. Dans un premier temps, nous délimitons à partir des données rassemblées et décrivons notre corpus de travail à l’aide de différentes mesures, en faisant référence aux nombreux travaux existants. Dans un deuxième temps, nous construisons la liste des têtes de segments et nous rappelons les apports de la littérature sur les noms sous-spécifiés. Partant des constructions spécificationnelles dans lesquelles ils s’inscrivent généralement, nous produisons une liste d’indices pour caractériser nos têtes de segments. Nous utilisons ces caractéristiques pour essayer de montrer en quoi ces têtes se rapprochent des emplois en noms sous-spécifiés en nous appuyant notamment sur leur forte fréquence et leur transdisciplinarité. Dans un troisième temps, nous discutons des schémas récurrents et fréquents dans lesquels s’insèrent les têtes de segments, qu’ils soient intra-segmentaux ou inter-segmentaux. Nous voulons caractériser de façon détaillée ces schémas au niveau syntaxique et sémantique, que cela soit pour le contenu ou le fonctionnement discursif. Nous voulons également étudier ces schémas à la lumière des constructions spécificationnelles bien identifiées par nos prédécesseurs. Enfin, dans un quatrième temps, nous discutons nos résultats et ouvrons de nouvelles perspectives.

# I. Exploration du corpus à la lumière de l’état de l’art

## I.1 Origine des données et prétraitement des données

### I.1.1 Récupération des données

L’accès aux titres a été grandement facilité par la création de bases de données bibliographiques, dont celles des archives ouvertes. Chaque chercheur, quelle que soit sa discipline, ou documentaliste d’un centre de recherche, est libre de déposer un document sur HAL avec l’accord des auteurs. Une archive ouverte présente l’avantage de centraliser l’accès aux travaux scientifiques, d’aider à leur diffusion et de les conserver manière pérenne, par rapport au site d’une institution particulière ou le site web personnel d’un chercheur, et de façon gratuite et accessible à tous, au contraire des éditeurs.

Nous utilisons le corpus constitué par Tanguy et Rebeyrolle (à paraître) comprenant près de 340  000 titres. Pour obtenir une si grande quantité de titres français, ils se sont tournés vers l’archive ouverte Hyper Article en Ligne (HAL, <https://hal.archives-ouvertes.fr>) (Nivard, 2010). Cette archive fonctionne depuis 2001 et est gérée par le Centre pour la Communication Scientifique directe du Centre National pour la Recherche Scientifique (CNRS). Elle contient plus de 1,6 millions de références, soit de travaux dont elle possède une copie, soit par le biais d’une notice. Plusieurs institutions, dont le CNRS, encourage le dépôt sur HAL des travaux produits par leurs chercheurs, garantissant un nombre important de titres issus de plusieurs disciplines. Alors que la majorité de la littérature traite des titres en anglais, HAL permet d’avoir accès à un grand corpus de titres en français. Nous veillerons dans ce premier chapitre à vérifier sur notre corpus certains enseignements tirés de l’étude des titres en anglais.

Chaque titre est fourni avec cinq informations supplémentaires relatives à la publication titrée :

1. un **identifiant** unique de la publication et donc du titre
2. les prénoms et noms des **auteurs** de la publication dont on peut déduire le nombre d’auteurs,
3. le **type** du document qui peut-être un article scientifique, un chapitre d’un ouvrage collectif ou une communication dans un congrès ou une conférence,
4. l’**année** de publication,
5. les **domaines scientifiques**, ou disciplines académiques, auxquels est associée la publication dont nous déduisons un domaine principal selon la méthode de Tanguy et Rebeyrolle (à paraître).

L’exemple n°2 ci-dessous montre les différentes informations pour un titre donné :

(2) Titre n°609897 : Villes durables et changement climatique : quelques enjeux sur le renouvellement des ressources urbaines ; 2011 ; ART ; 0.sde,1.sde.mcg ; Véronique Peyrache-Gadeau, Bernard Pecqueur.

HAL possède de nombreux types de documents différents. La majorité de la littérature traitant des titres d’articles de journaux scientifiques, notre corpus se limite à ceux-ci et à ceux construits de manière similaire : chapitres d’ouvrages collectifs et communications dans des conférences.

HAL permet d’attribuer plusieurs domaines à un document. Les domaines sont organisés en un arbre possédant quatre niveaux de profondeur, néanmoins la granularité des branches est très variable : « Sciences de l'Homme et Société » est une des racines de l’arbre, regroupant de nombreuses disciplines scientifiques, tout comme « Science non linéaire » et « Économie et finance quantitative ». Tanguy et Rebeyrolle (à paraître) propose une méthode de recodage des domaines pour n’en garder qu’un seul, le plus important et discriminant, que nous utilisons. Dorénavant, un titre est associé à un seul domaine principal.

### I.1.2 Étiquetage et analyse syntaxique en dépendances

Les titres ont été analysé à l’aide du logiciel Talismane (Urieli et Tanguy, 2013 ; Urieli, 2013) qui fournit un découpage en différents éléments, mots et signes de ponctuation, et fait un étiquetage morphosyntaxique des mots et une analyse syntaxique en dépendances des éléments. Pour chaque élément du titre nous avons :

* sa **forme** dans le titre,
* son **lemme** (pour les mots),
* sa **classe grammaticale/catégorie** (pour les mots, sinon nous avons “signe de ponctuation”)
* des **informations complémentaires**
* son élément **régisseur**,
* le **type de dépendance** qui le lie à son régisseur.

Les informations complémentaires dépendent de la classe grammaticale, comme le genre pour les noms, le mode et le temps pour les verbes. Les titres étant des textes très travaillés, ils ne nécessitent pas de prétraitement pour corriger les fautes, même s’il y en a de très rares comme l’oubli d’un point (3) ou le redoublement d’une préposition (4) :

(3) Titre n°62434 Développement stratégique du tourisme sportif de rivière par régulation corporatiste **L'**expérience du bassin de Saint Anne (Québec) appliquée aux Rivières de Provence

(4) Titre n°1559698 : Dispositif **de de** caractérisation simultanée de l'abondance de pucerons et de la croissance végétative d'arbres fruitiers

Il est à noter que Talismane a été conçu pour analyser des textes beaucoup plus longs que des titres et entraîné sur de tels textes. On peut donc douter de sa capacité à analyser correctement les titres. Notamment, comme nous le verrons plus tard, les titres ne comportent souvent pas de verbes conjugués au contraire des phrases de textes plus longs, ce qui pourrait pousser Talismane à reconnaître comme verbe des mots n’en étant pas. Nous avons décidé d’inclure une phrase de vérification de l’analyse de Taslimane lors de l’étape de sélection des racines.

### I.1.3 Segmentation des titres

Nous avons segmenté les titres selon la liste des signes de ponctuation segmentant établie par Anthony (2001). Nous en retranchons le tiret car il est utilisé pour lier de nombreux mots en français. Nous y ajoutons le point d’exclamation et les points de suspension. Nous avons donc les signes segmentant suivant :

|  |  |
| --- | --- |
| Ponctuation forte | . ? ! ... |
| Ponctuation faible | ; : |

Il y a dans cette liste des signes de ponctuation forte, comme le point ou le point d’interrogation, et des signes de ponctuation faible comme le point-virgule ou le double-point. Nous qualifions de segmentation forte une segmentation reposant sur une ponctuation forte et de segmentation faible une segmentation reposant sur une ponctuation faible.

L’analyse syntaxique en dépendances effectuées par Talismane ne va pas se comporter pareillement selon le type de segmentation. Une segmentation forte produit en effet deux phrases alors qu’une segmentation faible ne produit qu’une seule phrase. Talismane est bien plus capable de reconnaître une racine dans un segment qui est une phrase que dans un segment qui est une partie de phrase (5) même s’il n’en est pas toujours capable (6) :

(5) Titre n°760329 : L'**omniprésence** de la famille au sein de l'exploitation agricole : une *situation* de fait encouragé par les règles de droit

(6) Titre n°216312: **MODÈLES** THÉOTIQUES DE LA STRUCTURE DES JOINTS DE GRAINS.LES *MODÈLES* DE STRUCTURE DES JOINTS DE GRAINS ET LEUR UTILISATION

Dans les deux exemples précédents, omniprésence et modèles (en gras) sont bien reconnus comme des têtes de segments mais pas situation et modèles (en italique). En sortie de l’analyse en dépendances de Talismane, nous aurons des segments avec ou sans tête, voir plusieurs têtes en cas de comportements anormals de l’analyse qui est censée produire un arbre avec une racine unique : dans le cas d’un titre respectant le schéma “NC1 et NC2” par exemple, la racine choisit par Talismane est NC1, et dépend de NC1 et NC2 dépend de et.

### I.1.4 Sélection de la racine des segments

Une fois les segments délimités, il nous faut trouver leur racine. Pour les trouver et les compter, deux solutions s’offraient à nous. La première est une règle qui consiste à prendre le verbe conjugué du segment comme tête s’il y en a un, sinon une préposition si elle occupe la première position du segment et sinon le premier nom rencontré. Cette solution présente l’avantage d’être très simple mais nous avions peur de manquer des phénomènes remarquables ou de sélectionner le mauvais mot comme tête en nous basant si fortement sur la position.

Nous avons donc opté pour la seconde solution qui consiste à utiliser l’outil Talismane pour effecter une analyse syntaxique en dépendances. Cette analyse doit ramener pour chaque segment une tête correctement identifiée. Il s’agit d’une utilisation “a minima” de l’analyse en dépendances pour faire émerger une tête mais cela n’a pas été sans problème. La segmentation que nous effectuons, basé sur des signes de ponctuation segmentants, est décorrélée de l’analyse de l’outil, nous obtenons donc des segments sans tête. En prévision du chapitre suivant, nous nous limitons aux titres avec au maximum deux segments. On peut classer nos résultats en trois structures segments-racines :

1. Des titres ayant un segment et une racine
2. Des titres ayant deux segments dont un seul a une racine (soit le premier, soit le second)
3. Des titres ayant deux segments avec une racine dans chaque

La fiabilité de Talismane n’étant pas assurée sur des énoncés courts et généralement averbaux comme des titres, nous avons décidé d’estimer sa fiabilité. Nous avons choisi un échantillon de 20 titres aléatoirement pour chaque structure, en différenciant le cas deux selon que le segment sans racine est le premier et le second. Nous avons également choisi 20 titres ayant un segment et deux racines pour observer cet ensemble et éventuellement tenter d’en reprendre des titres. Nous avons vérifié manuellement le choix de la racine, sa catégorisation morphosyntaxique et son lemme. Les résultats complets sont dans l’annexe A3.B. Si globalement, Talismane arrive à étiqueter morphosyntaxiquement et à trouver le lemme correctement dans des énoncés aussi courts que des titres, la fiabilité pour sélectionnner la racine diffère grandement selon la structure segments-racines.

Avant d’aborder les résultats structure par structure, un premier point émerge : Talismane ne catégorise comme relation de dépendance racine (“root”) que les verbes. Pour les autres catégories, il reconnaît que la tête est l’élément racine de l’arbre de l’analyse en dépendances mais sans la qualifier de racine (“\_”). Le second point qui émerge concerne les segments sans racine dans les titres ayant deux segments : on constate l’existence d’un mot qui est régi par un mot de l’autre segment, généralement sa racine. D’après nos analyses manuelles, ce mot est la tête de l’autre segment. Nous avons donc développé un algorithme de sélection des têtes pour suppléer les déficiences de Talismane tout en gardant le bénéfice de l’analyse syntaxique en dépendances. Notre algorithme est présenté en détail après les résultats.

#### A. Titres avec un segment et une racine

Sur les 20 titres pris, Talismane a à chaque fois détecté la bonne racine, avec la bonne catégorie morphosyntaxique et le bon lemme, sauf une fois, où l'absence d'un accent ne lui a pas permi de reconnaître de lemme. On peut donc estimer que les titres qui suivent cette structure sont correctement analysés par Talismane.

#### B. Titres avec un segment et deux racines

Sur les 20 titres pris, Talismane a analysé incorrectement 12 titres et 8 ont une analyse discutable. Nous ne considérons pas le tiret et la virgule comme des caractères segmentants alors qu'ils sont clairement utilisés comme tels par un titre pour le tiret et deux titres pour la virgule. De plus, les mots composés des erreurs d'analyse dans Talismane qui désigne comme tête la partie après le tiret. Enfin, on remarque un oubli de signe de ponctuation segmentant et un crochet droit utilisé comme signe de ponctuation segmentant.

Nous pourrions changer notre liste de caractères segmentants, mais cela reviendrait à créer potentiellement de nouvelles erreurs. Nous décidons donc de ne pas utiliser les titres ayant deux têtes dans un seul segment.

#### C. Titres avec un segment ayant une racine suivie d’un segment sans racine

Sur les 20 titres, notre algorithme permet de sélectionner une tête valide dans le segment n'en contenant pas pour 17 d'entre eux. Deux titres utilisent la virgule comme un caractère segmentant. Enfin un dernier échappe à notre algorithme de sélection.

#### D. Titres avec un segment sans racine suivi d’un segment avec racine

Sur les 20 titres, notre algorithme permet de sélectionner une tête valide dans le segment n'en contenant pas pour 18 d'entre eux. On note des erreurs d'analyse de Talismane liées à une mauvaise catégorisation morphosyntaxique de mots dont cinq entraîne une mauvaise sélection de la tête.

#### E. Titres avec un segment avec racine suivi d’un segment avec racine

Sur les 20 titres, 16 sont correctement analysés par Talismane qui trouve les têtes des segments. Pour trois titres la tête est mal catégorisée et pour un dernier le lemme n'est pas trouvé.

#### F. Algorithme de sélection de tête de segment

Notre algorithme pour détecter la tête d’un segment lorsque Talismane n’y arrive pas est le suivant :

On regarde si un mot du segment sans racine est régi par la racine de l'autre segment :

Si oui ⇒ on prend le premier mot correspondant à cette définition comme tête

Si non ⇒ on regarde si le premier mot du segment sans racine est régi par un mot de  
 l'autre segment :

Si oui ⇒ on le prend comme tête

Si non ⇒ impossible de trouver une tête

Une fois les données récupérées et prétraitées, nous constituons notre corpus de travail. Il faut pour cela établir un périmètre qui délimitera notre corpus de travail. Il faut expliquer le choix de notre périmètre et effectuer des mesures dessus, afin de mettre en relation notre corpus de travail avec ceux étudiés précédemment dans la littérature.

## I.2 Description des données et mesures du corpus

### I.2.1 Description des données des titres

Nous avons comme données un ensemble de 339 687 titres ayant les caractéristiques suivantes :

* identifiant,
* année,
* type de support (article, chapitre ou communication),
* domaine,
* auteurs,
* nombre d’auteurs,
* énoncé,
* liste de mots et de signes de ponctuation que nous appelons “éléments du titre” :
  + Pour chaque élément :
    - forme
    - étiquette morphosyntaxique
    - lemme (toujours égale à sa forme pour un signe de ponctuation)
    - informations supplémentaires
    - élément régisseur
    - type de relation de dépendance
    - sa position dans le titre
* longueur du titre en nombre d’éléments (mots + signes de ponctuation),
* longueur du titre en nombre de mots (mots uniquement),
* segments :
  + Permet d’accéder aux différents segments du titre et notamment :
    - tête du syntagme,
    - sa position dans le titre,
* nombre de segments.

On notera que les différentes données ne sont pas indépendantes : Kutch (1978), Yitzhaki (1994) et Tanguy et Rebeyrolle (à paraître) ont ainsi montré que le nombre d’auteurs et corrélé positivement à la longueur du titre. Larivière et al. (2015) que le domaine est lié au nombre d’auteurs : il y a en moyenne plus d’auteurs dans les sciences exactes. Baethge (2008) a montré que le nombre d’auteurs augmente avec le temps. Après avoir décrit nos données nous établissons le périmètre qui délimitera notre corpus de travail. Tanguy et Rebeyrolle (à paraître) ont également montré, en partant des mêmes données de base et donc avec le même déséquilibre de répartition, que la longueur était très légèrement corrélée à l’année de publication.

### I.2.2 Sélection des données selon la structure et donc la nature des titres

Dans un premier temps, on effectue le choix du périmètre de la sélection des données qui vont constituer notre corpus de travail. Ce périmètre est défini par rapport à la structure des titres et la nature de la tête des segments du titre.

#### A. Structures des titres

Nous avons décidé de prendre les titres composés de 1 ou 2 segments. Nous justifions ce choix par le fait qu’il s’agit de la plus grande majorité des titres (320 561 soit 94 % des titres initiaux) et qu’ils sont plus faciles à analyser. De nombreux travaux didactiques sur l’écriture des titres (Aleixandre-Benavent et al., 2014 ; Swales et Feak, 1994 ; Gustavii, 2008) conseillent d’organiser les titres en deux segments autour d’un double point : “*A : B*”.

Un autre délimiteur que nous utilisons pour établir notre périmètre, en plus du nombre de segments dans le titre, et le nombre de têtes par segments. Nous nous limiterons aux titres avec au maximum une tête par segment. On distingue donc deux cas : les titres composé d’un seul segment et les titres composés de deux segments.

##### A.1 Titres composés d’un seul segment

Exemples de titres :

(7) Titre n°360059 :  
1 2 3 4 5 6 7 8  
L' **actualité** de la jurisprudence communautaire et internationale  
DET NC P DET NC ADJ CC ADJ  
2 0 2 5 3 5 6 7  
det \_ dep det prep mod coord dep\_coord

(8) Titre n°1258610 :  
1 2 3 4 5   
**Doit** -on écouter Björk ?  
V CLS VINF NPP PONCT  
0 1 1 3 4  
root suj obj obj ponct

Il y a 171 890 titres composé d’un seul segment ayant une seule tête de segment, soit près de 51% des titres récupérés initialement.

##### A.2 Titres composés de deux segments

Exemples de titres :

(9) Titre n°1258625 :  
1 2 3 4 5 6 7 8 9 10 11 12  
Un nouvel **OVNI** dans le ciel réunionnais : la **transparence** des prix  
DET ADJ NC P DET NC ADJ PONCT DET NC P+D NC  
3 3 0 3 6 4 6 7 10 0 10 11  
det mod \_ dep det prep mod ponct det \_ dep prep

(10) Titre n°360068 :  
1 2 3 4 5 6 7 8 9 10  
La **performativité** de l' évidence : analyse du discours néolibéral  
DET NC P DET NC PONCT NC P+D NC ADJ  
2 0 2 5 3 5 2 7 8 9  
det \_ dep det prep ponct mod dep prep mod

Il y a 124 938 titres composé de deux segments, soit près de 37 % des titres récupérés intialement. Mais, du fait des limites entre les capacités de Talismane et notre définition des segments, certains segments n’ont pas de tête. Nous avons créé un algorithme pour suppléer ces limitations. Si dans les exemples précédents, le titre n°1258625 a bien deux segments avec une racine par segment, le titre n°360068 a un segment, son second, sans racine. Notre algorithme promeut “analyse” comme tête du segment car ce mot est régi par la tête de l’autre segment. Lorsqu’aucun mot du segment n’ayant pas de racine n’est régi par la tête de l’autre segment, nous regardons si le premier mot du segment sans tête est régi par un mot, n’importe lequel, du segment ayant une tête et si oui, nous prenons le mot régi comme tête du segment qui n’en avait pas.

Nous arrivons ainsi, sur les 56 851 titres ayant 2 segments mais une seule racine, à récupérer 46 798 titres soit 82 % d’entre eux. Pour finir nous gardons 110 785 titres avec deux segments avec une tête dans chaque. Nous avons donc 171 890 titres monosegmentaux (61 %), 110 785 bisegmentaux (39 %), soit un corpus de travail de 282 675 titres, ce qui représente 83 % du matériau de base, les presque 340 000 titres collectés sur HAL. Un corpus de titres ne sera toujours qu’un sous-ensemble de l’infinité des énoncés possibles pour la fonction de titre. Garder un corpus de travail d’une taille importante permet d’essayer de tendre vers la réalité de l’ensemble des énoncés produits, sans jamais y arriver, et renforcer la crédibilité des hypothèses que l’on teste sur lui ou que l’on émet à partir de son observation. Nous avons réussi à conserver 83 % du matériau initial dans cette première étape de définition du périmètre de notre corpus de travail, néanmoins nous restreignons encore notre périmètre dans l’étape suivante pour nous intéresser à une catégorie morphosyntaxique particulière.

#### B. Nature des titres

Nous nous sommes interrogés sur la nature de la tête des segments pour opérer une sélection sur ce critère. Cette question est directement liée à la question de la nature des titres. D’après Schwischay (2001), « un nœud forme avec tous les nœuds qu’il domine (directement ou indirectement) un syntagme ; et, par convention, ce syntagme porte le nom du nœud dominant ». Nous pouvons donc, grâce à la complémentarité du modèle de l’analyse en constituants immédiats et celui de l’analyse en dépendances, déterminer le type de syntagme de chaque segment en étudiant la catégorie morphosyntaxique de sa tête.

|  |  |  |  |
| --- | --- | --- | --- |
| **Catégorie morphosyntaxique** | **Titres monosegmentaux** | **Titres bisegmentaux, segment 1** | **Titres bisegmentaux, segment 2** |
| **Noms communs** | 136 734 (80 %) | 82 959 (75 %) | 84 960 ( 77 %) |
| **Noms propres** | 11 094 ( 6 %) | 10 406 ( 9 %) | 4 758 ( 4 %) |
| **Noms** | **147 828 (86 %)** | **93 365 (84 %)** | **89 718 ( 81 %)** |
| **Verbes à l’indicatif** | 8 186 ( 5 %) | 3 478 ( 3 %) | 3 513 ( 3 %) |
| **Verbes à l’infinitif** | 5 135 ( 3 %) | 6 004 ( 5 %) | 2 140 ( 2 %) |
| **Verbes** | **15 749 ( 9 %)** | **10 672 (10 %)** | **6 549 ( 6 %)** |
| **Prépositions** | **6 792 ( 4 %)** | **5 456 ( 5 %)** | **10 456 ( 9 %)** |

Distribution des catégories morphosyntaxiques des têtes de segment

On peut remarquer des points communs : la grande majorité des têtes sont des noms, et a fortiori des noms communs, pour toutes les configurations segmentales. Les autres catégories les plus représentées sont les verbes à l’indicatif ou à l’infinitif et les prépositions. La différence la plus notable entre les premiers et seconds segments des titres bisegmentaux et que pour les seconds segments, la seconde catégorie la plus fréquente sont les prépositions et non les verbes : les têtes prépositionnelles sont presque deux fois plus fréquentes que dans les segments des titres monosegmentaux et dans les premiers segments des titres bisegmentaux.

On peut ensuite s’interroger sur les combinaisons possibles dans les titres bisegmentaux entre les catégories des deux têtes de segments. Le tableau suivant présente les combinaisons les plus fréquentes couvrant 90 % des titres bisegmentaux :

|  |  |  |
| --- | --- | --- |
| **Catégorie de la tête du premier segment** | **Catégorie de la tête du second segment** | **Nombre de titres et pourcentage** |
| NC | NC | 63 719 ( 56 % ) |
| NPP | NC | 7 974 ( 7 % ) |
| NC | P | 6 763 ( 6 % ) |
| VINF | NC | 4 643 ( 4 % ) |
| NC | NPP | 3 420 ( 3 % ) |
| P | NC | 3 283 ( 3 % ) |
| NC | V | 2 565 ( 2 % ) |
| V | NC | 2 534 ( 2 % ) |
| NC | VINF | 1 486 ( 1 % ) |
| NC | P+D | 1 295 ( 1 % ) |
| NC | CC | 1 168 ( 1 % ) |
| P+D | NC | 1 030 ( 1 % ) |

Le tableau suivant aggrège les différentes catégories nominales, verbales et prépositionnelles en trois catégories : Nom, Verbe et Préposition.

|  |  |  |
| --- | --- | --- |
| **Catégorie de la tête du premier segment** | **Catégorie de la tête du second segment** | **Nombre de titres et pourcentage** |
| Nom | Nom | 75 592 ( 68 % ) |
| Nom | Préposition | 8 996 ( 8 % ) |
| Verbe | Nom | 8 506 ( 8 % ) |
| Nom | Verbe | 5 426 ( 5 % ) |
| Préposition | Nom | 4 650 ( 4 % ) |

Pour les titres monosegmentaux, déterminer la nature du titre revient à prendre la nature de son unique segment. On obtient donc que 86 % de titres nominaux. Pour les titres bisegmentaux, on peut considérer qu’un titre est nominal si son premier segment l’est. On obtient alors 84 % de titres nominaux. Une autre solution est de considérer qu’un titre est “purement” nominal si et seulement si les deux têtes de ses segments sont des noms. On obtient alors 68 % de titres nominaux.

Quelle que soit la solution choisie, les titres sont majoritairement constitués d’un ou plusieurs syntagmes nominaux, ce qui rejoint les conclusions de nos prédécesseurs (Leech, 2000 ; Haggan, 2004 ; Soler, 2007 ; Cheng et al., 2012 ; Wang et Bai, 2007), jusqu’à 93 % de titres nominaux pour le corpus de Cheng et al. (2012) et 99 % pour celui de Wang et Bai (2007), et non d’une phrase avec un noyau verbal.

Pour notre corpus de travail, nous décidons de nous restreindre aux titres monosegmentaux dont la tête est un nom et aux titres bisegmentaux dont au moins une des têtes de ses segments est un nom, l’autre pouvant être un nom, une préposition ou un verbe. Ce choix nous permet de garder la majorité de nos titres et d’éliminer les cas peu fréquents.

Une fois le périmètre des titres étudiés défini sur la structure segmentale des titres et la nature grammaticale de leurs têtes, nous avons constitué notre corpus de travail. Nous pouvons alors effectuer plusieurs mesures dessus et les mettre en rapport avec les mêmes mesures effectuées dans des travaux précédents, avant d’étudier plus avant les têtes de syntagmes.

### I.2.3 Mesures effectuées sur notre corpus de travail

Nous avons défini notre périmètre d’étude comme portant sur les titres constitués d’un ou deux segments. Les titres monosegmentaux (147 828, 59 %) ont une tête nominale, les titres bisegmentaux (103 170, 41 %) ont un segment ayant une tête nominale, l’autre ayant une tête verbale, nominale ou prépositionnelle. Nous obtenons un corpus de 250 998 titres, soit 74 % du matériau initial.

Sur la longueur des titres, les titres monosegmentaux ont une longueur moyenne de 10,38 mots, avec une longueur minimum de 1 mot et un maximum de 77 mots, tandis que les titres bisegmentaux ont une longueur moyenne de 14,45 mots, avec une longueur minimum de 2 mots et une longueur maximum de 228 mots. Les titres bisegmentaux les plus courts sont au nombre de 64, 49 utilisent comme signe segmentateur le double point et 51 sont des chapitres d’ouvrage dont 29 sont de la forme Entrée : NC, indiquant une entrée dans un ouvrage de type dictionnaire ou encyclopédie. La longueur supérieure des titres bisegmentaux s’explique par la facilité de traitement qu’apporte la segmentation à l’interlocuteur : la segmentation sert à la fois de pause et d’articulation pour sa compréhension. La longueur moyenne des titres du corpus de travail est de 12,05 mots, alors que celle des données de départ est de 13,8 mots. C’est normal car il existe des titres ayant plus de deux segments que notre corpus de travail n’inclut pas.

On peut regarder comme nos corpus se répartit en fonction du type de publication scientifique :

|  |  |  |  |
| --- | --- | --- | --- |
| **Type de publication** | **Titres monoseg.** | **Titres biseg.** | **Corpus** |
| Article | 63 993 43 % | 45 827 44 % | 109 820 44 % |
| Communication | 53 148 36 % | 35 350 34 % | 88 498 35 % |
| Chapitre d’ouvrage | 29 413 20 % | 21 221 21 % | 50 634 20 % |
| Poster | 1 274 1 % | 772 1 % | 2 046 1 % |

La structure des titres n’est pas corrélée au type de publication, la distribution des deux ensembles étant presque identique. De plus, cette répartition est quasi identique à celle de l’ensemble des 340 000 titres qui constituent nos données de départ (Tanguy et Rebeyrolle, à paraître).

On peut aussi mesurer le nombre d’auteurs en fonction de la structure du titre :

|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre d’auteurs** | **Titres monoseg.** | **Titres biseg.** | **Corpus** |
| 1 | 87 646 59 % | 65 199 63 % | 152 845 61 % |
| 1-4 | 135 564 92 % | 96 581 94 % | 232 145 92 % |
| 1-9 | 146 767 99 % | 102 307 99 % | 249 074 99 % |

On voit bien que quelle que soit sa structure, la répartition par le nombre d’auteurs et la même pour les deux sous-ensembles de notre corpus de travail que pour le corpus de travail pris dans sa totalité, et sur l’ensemble des données, 62 % des articles avaient également un seul auteur.

On regarde également la répartition par années de publication. Pour l’ensemble du corpus, elles s’étendent de 2019 pour les sept publications les plus récentes à 1779 pour la plus ancienne. On note que 85 % des publications ont été publié en 2000 ou après, 90 % après 1994 et 99 % après 1933. Pour l’ensemble des données Tanguy et Rebeyrolle (à paraître), trouve les mêmes années pour les deux premiers pourcentages et un peu plus tard, 1940, pour le dernier. Notre corpus ne peut donc pas servir pour des études diachroniques du fait de sa répartition totalement inégale sur le temps. La période qui comporte le plus de titres, de 2005 à 2017, 74 % du corpus, est également trop courte. La répartition est similaire pour nos deux sous-corpus, titres monosegmentaux et bisegmentaux.

Nous regardons la répartition des titres par domaine à présent pour le corpus et les deux sous-corpus. Nous rappelons que nous avons sélectionné, grâce à la méthode décrite dans Tanguy et Rebeyrolle (à paraître) un seul domaine principal pour chaque titre. Le tableau suivant présente les 27 domaines qui existent dans notre corpus. Nous avons mis en gras les domaines des sciences exactes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **N°** | **Domaine** | **Corpus**  **Nb/fréq/fréq. cumul** | **Répartition entre** | |
| **Titres monosegmentaux** | **Titres bisegmentaux** |
| **01** | **Physique** | 26 559 11% 11% | **81 %** | 19 % |
| 02 | Sociologie | 23 732 9% 20% | 48 % | 52 % |
| 03 | Droit | 21 486 9% 29% | 67 % | 33 % |
| 04 | Histoire | 19 093 8% 36% | 54 % | 46 % |
| 05 | Pas de domaine associé | 18 941 8% 44% | 59 % | 41 % |
| 06 | Gestion et management | 18 318 7% 51% | 45 % | 55 % |
| **07** | **Sciences du vivant** | 17 498 7% 58% | **66 %** | 34 % |
| **08** | **Informatique** | 13 505 5% 63% | **74 %** | 26 % |
| **09** | Linguistique | 11 556 5% 68% | 52 % | 48 % |
| 10 | Littératures | 10 712 4% 72% | 52 % | 48 % |
| 11 | Archéologie et Préhistoire | 10 124 4% 76% | 61 % | 39 % |
| 12 | Science politique | 7 152 3% 79% | 46 % | 54 % |
| 13 | Éducation | 7 062 3% 82% | 50 % | 50 % |
| 14 | Art et histoire de l'art | 6 471 3% 85% | 53 % | 47 % |
| 15 | Philosophie | 6 152 2% 87% | 60 % | 40 % |
| 16 | **Sciences de l'environnement** | 5 542 2% 89% | 54 % | 46 % |
| 17 | Sciences de l'information et de la communication | 5 481 2% 91% | 46 % | 54 % |
| 18 | Anthropologie | 5 166 2% 93% | 51 % | 49 % |
| 19 | Architecture | 3 444 1% 95% | 51 % | 49 % |
| 20 | **Planète et Univers** | 2 781 1% 96% | **62 %** | 38 % |
| 21 | **Mathématiques** | 2 377 1% 97% | **81 %** | 19 % |
| 22 | **Sciences cognitives** | 2 370 1% 98% | 53 % | 47 % |
| 23 | **Chimie** | 2 185 1% 99% | **69 %** | 31 % |
| 24 | Psychologie | 2 006 1% 99% | 54 % | 46 % |
| 25 | Géographie | 860 0% 100% | 51 % | 49 % |
| 26 | **Économie et finance quantitative** | 346 0% 100% | 47 % | 53 % |
| 27 | Autres | 79 0% 100% | 54 % | 46 % |
|  | **Sciences exactes** | 73 163 29% | 72 % | 28 % |
| moyenne 65 % écart-type 0.11  écart-type relatif 18 % | |
|  | Sciences humaines et sociales | 177 835 71% | 54 % | 46 % |
| moyenne 53 %  écart-type 0.06  écart-type relatif 10 % | |

On compte 73 163 titres en sciences exactes, ce qui représente 29 % de notre corpus et 177 835 titres en sciences humaines et sociales, soit 71 %.

Les sciences exactes globalement privilégient plus les titres monosegmentaux que les sciences humaines et sociales. Si l’on regarde la moyenne des répartitions par domaine, l’écart-type relatif important nous pousse néanmoins à la prudence. Parmi les sciences exactes, les mathématiques et la physique utilisent le plus fréquemment des titres monosegmentaux, où ils représentent 81 % des titres. Ces domaines sont suivis par l’informatique où ils représentent 74 % des titres, suivie de la chimie avec 69 %, des sciences du vivant avec 66 % et des sciences des planètes et de l’univers avec 62 %.

Les sciences humaines et sociales sont globalement plus équilibrées dans l’utilisation de titres monosegmentaux et bisegmentaux. Mais l’écart-type relatif de 10 % montre que cet équilibre global varie d’un domaine à l’autre. Ainsi le droit avec 67 %, l’archéologie et la préhistoire avec 61 % et la philosophie avec 60 % privilégient elles aussi le titre monosegmental.

Si on compare la répartition par domaine de notre corpus de travail, nous avons le même ordre que celui relevé par Tanguy et Rebeyrolle (à paraître) pour l’ensemble des données initiales. Nous notons également que la répartition entre les domaines n’est pas homogène, certains étant très peu représentés, les plus faiblement dotés étant la géographie avec 860 titres, l’économie et finance quantitative avec 346 titres, et autres avec 79 titres. D’où la nécessité de travailler en fréquence relative pour les phénomènes que nous étudierons tout en retenant qu’une fréquence relative peut dissimuler un très petit phénomène : un phénomène ayant une fréquence relative de 15 % dans le domaine autre, ne concernera finalement que 11 titres.

Nous avons dans ce sous-chapitre établi le périmètre délimitant notre corpus de travail et mesuré ses contours. Nous avons décidé d’étudier le cas le plus nombreux : celui des titres mono ou bisegmentaux possédant au moins une tête nominale. Cela nous amène à vouloir étudier les têtes de ces segments pour éventuellement les rapprocher d’un emploi particulier des noms généraux : les noms sous-spécifiés.

# II. Caractérisation des têtes de segments

Dans cette partie, nous nous intéressons aux têtes de nos segments. Nous avons vu que nous avons une tête par segment et d’un à deux segments par titre. Cela fait donc trois sous-ensembles de notre corpus de travail : les segments des titres monosegmentaux, les premiers segments des titres bisegmentaux et les seconds segments des titres bisegmentaux. Nous allons étudier dans ces trois ensembles les têtes de segments. De plus, nous gardons à l’esprit que nous voulons étudier la variété disciplinaire : il peut donc être possible de mettre avant des têtes très représentatives d’une discipline. Au contraire, nous pouvons aussi circonscrire un ensemble de têtes très fréquentes dans de nombreux domaines, des têtes que nous appellerons transdisciplinaires. Ce sont ces dernières têtes, que nous voulons rapprocher des noms généraux sous-spécifiés ensuite. Les têtes transdisciplinaires sont en effet très fréquentes et leur capacité à être présentes dans de nombreux domaines indique un sens général, et donc un contenu sémantique pauvre, qui appelle à être spécifié, ou “rempli”, par le contexte. Cette opération de remplissage s’effectue pour les noms généraux sous-spécifiés par des constructions spécificationnelles dans lesquelles le nom à spécifier et le contexte spécifiant s’inscrivent. On peut donc se demander si on retrouve ces constructions étudiées par la littérature dans nos titres et si, en leur absence, ne pouvons-nous pas en mettre à jour de nouvelles.

## II.1 Têtes de segments

Nous avons 27 domaines et sur l’ensemble de notre corpus de travail et 354 168 occurrences de têtes, 1 pour chacun des 147 828 titres monosegmentaux, et 2 pour chacun des 103 170 titres bisegmentaux. Nous pouvons regrouper les occurrences d’une même tête sous son lemme. Nous comptons 123 227 lemmes différents. Nous notons **OCCTÊTE** toutes les occurrences d’un lemme, noté **TÊTE**. Pour chaque **TÊTE**, nous avons :

* un nombre total d’occurrences dans le corpus, noté **NBCORPUS(OCCTÊTE)**,
* un nombre d’occurrences par domaine, noté **NBDOMAINE(OCCTÊTE)**,
* on note que la somme des **NBDOMAINE(OCCTÊTE)** = **NBCORPUS(OCCTÊTE)**,

Pour un domaine donné, on a également :

* un nombre total d’occurrences de têtes, noté **NBDOMAINE(OCC)**,
* un nombre de têtes différentes, noté **NBDOMAINE(TÊTE)**
* on note le nombre de domaines **NBDOM**

Soit le domaine de la physique. Il compte 2 471 occurrences de la tête *étude* qui compte 6 842 occurrences en tout dans le corpus. Le domaine physique compte 3 584 têtes différences, et 30 667 occurrences de têtes en tout. On a donc pour la tête *étude* les nombres suivants :

* **NBCORPUS(OCCÉTUDE)** = 6 842
* **NBPHYSIQUE(OCCÉTUDE)** = 2 471
* **NBPHYSIQUE(OCC)** = 30 667
* **NBPHYSIQUE(TÊTE)** = 3 584

Pour chaque tête, on peut donc établir deux séries statistiques : l’une est constitué des fréquences de la tête dans les différents domaines par rapport au nombre total de têtes dans ce domaine. Pour l’obtenir il faut diviser **NBDOMAINE(OCCTÊTE)** / **NBDOMAINE(OCC)**. Ce calcul représente le poids de la tête dans un domaine particulier : plus il est haut, plus la tête est utilisée de façon importante dans ce domaine. On notera la variable **FREQDOMAINE(OCCTÊTE)** et la série pour une tête donnée **FREQ(OCCTÊTE)**. La série pour un domaine donné des fréquences des occurrences de ses têtes s’appellera **FREQDOMAINE(OCC)**.

L’autre série repose sur la répartition du nombre total d’occurrences dans le corpus, **NBCORPUS(OCCTÊTE)**, dans les différents domaines, soit **NBDOMAINE(OCCTÊTE)** / **NBCORPUS(OCCTÊTE)**. Nous appelons cette série **DISTDOMAINE(OCCTÊTE).** La moyenne de cette série est égale à **NBCORPUS(OCCTÊTE)** divisé par le nombre de domaines, soit 27, ce qui correspondrait à une répartition égalitaire des occurrences entre les différents domaines. Cette hypothèse est intuitivement rejetée, mais nous pouvons essayer de mesurer l’ampleur de la différence entre la répartition réelle et la répartition égalitaire.

Avec ces chiffres à notre disposition, nous pouvons essayer de résoudre deux questions : quelles sont les têtes représentatives d’une discipline et quelles sont les têtes transdisciplinaires, des têtes que l’on retrouve fréquemment dans de nombreuses disciplines ?

### II.1.1 Les têtes de segments spécifiques

La question des têtes spécifiques à une discipline nous semble intéressant sur deux points. Le premier est de tester si nous pouvons réussir à faire émerger ces têtes et ainsi obtenir une liste de têtes spécifiques par domaine. Le second est de pouvoir la comparer à la liste de têtes transdisciplinaires que nous voulons également établir, pour si, entre têtes spécifiques et têtes transdisciplinaires, il s’agit d’une dichotomie ou d’un problème de seuil.

#### A. Définitions théorique et opératoire

Pour être véritablement spécifique à un domaine, une tête ne doit pas seulement y avoir beaucoup d’occurrences : ces occurrences doivent occuper une place importante dans le domaine considéré, la tête doit y être *fréquente*.

Nous commençons par sélectionner les têtes ayant une fréquence relative supérieure ou égale à 0,3 % dans le domaine donné, ce sera notre **seuil de fréquence**, par rapport au nombre de total de têtes dans ce domaine, soit **FREQDOMAINE(OCCTÊTE) >= 0.003**. La valeur du seuil a été déterminée de façon arbitraire après une série d’essais empiriques. Il s’agit d’un premier filtre pour ne garder que les têtes dont le nombre d’occurrences dans le domaine les rend assez fréquente pour être considérées.

Sur certains domaines où il y a très peu de titres et beaucoup de têtes différentes, ce minimum peut ne jamais être atteint. Ainsi si pour un domaine A donné, **MAX(FREQDOMAINE A(OCC)) < seuil de fréquence**,aucune tête ne sera sélectionnée. Au contraire, si le filtre est trop bas, toutes les têtes d’un domaine donné pourront être sélectionnées selon la formule **NBDOMAINE(OCC) \* seuil de fréquence < 1**, soit si **NBDOMAINE(OCC) < 1 / seuil de fréquence**.

Le seconde filtre que nous appliquons se base sur la différence entre l’hypothétique répartition égalitaire des occurrences d’une tête, soit **NBCORPUS(OCCÉTUDE)** / NBDOM, et la répartition réelle, **NBDOMAINE(OCCÉTUDE)** : nous prenons l’écart entre la répartition réelle et l’égalitaire que nous divisons par le nombre d’occurrence de la tête dans le corpus pour passer d’un nombre absolu à un pourcentage. On aura donc ( **NBDOMAINE(OCCÉTUDE) - NBCORPUS(OCCÉTUDE)** / NBDOM ) / **NBCORPUS(OCCÉTUDE)** qui se simplifie en **DISTDOMAINE(OCCTÊTE) - 1 /** NBDOM. Nous ne sélectionnons que les têtes pour lesquelles ce calcul dépasse ou égale notre **seuil de distribution**. Nous fixons celui-ci arbitrairement après une série d’essais empiriques à 0.025.

Notre filtre peut donc s’écrire avec le pseudo-code suivant :

**Filtre de sélection des têtes spécifiques à des domaines**

**pour chaque** DOMAINE **fait**  
 **si** NBDOMAINE(OCC) \* seuil de fréquence< 1 **alors**  
 on ne peut pas calculer les têtes spécifiques pour ce domaine  
 **fin si  
 Pour chaque** TÊTE **fait  
 si** FREQDOMAINE(OCCTÊTE) >= seuil de fréquence **alors**  
 **si** DISTDOMAINE(OCCTÊTE) - 1 / NBDOM >= seuil de distribution **alors**  
 **sélectionne** TÊTE  
 **fin si**  
 **fin si**  
 **fin pour  
fin pour**

##### Corrections de Talismane

Néanmoins, ce seuil demeure très faible dans l’absolu et rend notre filtre très sensible à un mauvais traitement d’un lemme par Talismane. Nous avons donc établi un dictionnaire des corrections pour essayer de corriger au maximum des erreurs de catégorisation et lemmatisation. Le tableau suivant liste certaines catégories d’erreurs. Pour savoir comment les corriger, nous avons regardé les différents titres concernés pour établir à chaque fois une règle ad-hoc :

|  |  |  |
| --- | --- | --- |
| **Erreur** | **Correction** | **Exemples** |
| Forme catégorisée comme nom propre avec un lemme inconnu car avec une majuscule. | Lemme ajouté, catégorie corrigée à nom commun. | Effet, Adolescence, Autoformation, Approche, Cohomologie |
| Forme non reconnue d’un nom commun car erreur d’orthographe | Lemme corrigé | Quanti**ﬁ**cation, év**è**nement,  indicateus (-**r**),  Synth**č**se |
| Forme non reconnue d’un nom commun car caractère non compris | Lemme corrigé (en écrivant oeuvre) | œuvre |
| Forme non connue d’un nom commun | Lemme ajouté | démotorisation, maritimisation, Compactification,  Ondelettes |
| Forme non reconnue d’un nom propre | Lemme corrigé | Paris, Freud |
| Forme faussement reconnue comme nom alors qu’il s’agit d’un adjectif | Forme non prise en compte et retirée de nos calculs | Cyber, Environnemental |
| E- et Semi- considérés comme un nom propre indépendant | Lemme corrigé en e- ou semi- + lemme suivant | E-chronic, E-commerce, E-administration, e-inclusion, Semi-figement |
| s considéré comme un nom commun à cause d’un signe de ponctuation | On regarde à gauche et à droit du s pour trouver un nom commun ou un nom propre après un signe de ponctuation | mobilité.s, Linguistique(s),  Quel(s) avenir(s) |
| Mot anglais non reconnu catégorisé à tort comme nom commun | Forme non prise en compte et retirée de nos calculs, provenant de titres en anglais. | The |
| Nom commun anglais non reconnu | Prise en considération de son lemme en français | Synthesis |
| Emploi d’un nom propre au pluriel | Lemme corrigé à la forme singulière | Venises |

Une fois ces corrections effectuées sur notre corpus de travail, nous pouvons passer notre filtre dessus pour obtenir les têtes spécifiques à certaines disciplines, en les classant par leur fréquence dans le domaine.

#### B. Résultats et évaluations des résultats

Il est possible de jouer sur les seuils de fréquence et de distribution pour restreindre les têtes sélectionnées au détriment de la richesse du résultat. Plus les seuils seront haut, moins il y aura de têtes spécifiques détectées. Seule l’évaluation des résultats permet de juger de la pertinence des valeurs des seuils et l’efficacité de notre méthode.

Nous avons filtré l’ensemble des têtes du corpus de travail et sélectionné 356 têtes spécifiques. Nous avons ignoré les titres sans domaine unique associé et le domaine autres car il est difficile de trouver des spécifiés à des domaines fourre-tout. De plus le domaine autres n’avait pas assez de titres pour que notre algorithme marche : dans notre cas si **NBDOMAINE(OCC)** doit être supérieur à 1 / 0.003, soit 333, ce qui disqualifie le domaine Autre qui a seulement 79 titres. Cela nous laisse 25 domaines auxquels nous essayons d’associer des têtes caractéristiques qui sont pris en compte dans nos calculs.

Nous proposons dans le tableau suivant un extrait, classé par domaine, des têtes spécifiques en prenant les dix premières selon l’ordre de leur **FREQDOMAINE(OCCTÊTE)**. La liste complète est fournie en annexe. Les noms propres sont en gras. Pour chaque domaine, classés par ordre alphabétique, nous indiquons quatre nombres : combien de lemmes de têtes ont été sélectionnées pour le domaine, le nombre de lemmes de têtes différents dans ce domaine, **NBDOMAINE(TÊTE)**, le nombre d’occurrences que représentent les lemmes sélectionnés en pourcentage par rapport au nombre de têtes dans ce domaine et cette dernière valeur, **NBDOMAINE(OCC)** :

|  |  |  |
| --- | --- | --- |
|  | **Domaine** | **Têtes associées** |
| 01 | Anthropologie  **7** / 2 579 / **4 %** / 6 942 | anthropologie, ethnographie, corps, mémoire, patrimoine, identité, objet |
| 02 | Archéologie et préhistoire  **34** / 3 444 / **21 %** / 13 391 | céramique, nécropole, sanctuaire, occupation, sépulture, site, dépôt, site, archéologie, décor, fouille |
| 03 | Architecture  **15** / 1 624 / **11 %** / 4 629 | ambiance, urbanisme, fortification, **Paris**, architecture, quartier, château, habitat, aménagement, ville |
| 04 | Art et histoire de l’art  **19** / 3 376 / **11 %** / 8 685 | vitrail, sculpture, artiste, peinture, cinéma, musique, collection, portrait, théâtre, décor |
| 05 | Chimie  **19** / 788 / **19 %** / 2 710 | ligand, hydrogénation, catalyse, catalyseur, membrane, oxydation, polymère, nanoparticule, réactivité, chimie |
| 06 | Droit  **38** / 4 189 / **23 %** / 26 398 | clause, obligation, juge, droit, assurance, chronique, contrat, responsabilité, commentaire, liberté |
| 07 | Économie et finance quantitative  **4** / 273 / **2 %** / 489 | aversion, **GRP**, complexification, tarification |
| 08 | Éducation  **37** / 1 786 / **24 %** / 9 445 | informatique, éducation, didactique, enseignant, pédagogie, accompagnement, orientation, école, formation, compétence |
| 09 | Géographie  **3** / 604 / **3 %** / 1 191 | démographie, écologie, migration |
| 10 | Gestion et management  **50** / 3 546 / **36 %** / 25 955 | management, déterminant, économie, entreprise, gouvernance, marché, innovation, crise, proposition, impact |
| 11 | Histoire  **19** / 7 005 / **9 %** / 25 671 | femme, mémoire, histoire, société, source, image, compte, ville, remarque, introduction |
| 12 | Informatique  **49** / 3 281 / **40 %** / 16 241 | ordonnancement, algorithme, segmentation, extraction, visualisation, planification, classification, plateforme, reconstruction, détection |
| 13 | Linguistique  **35** / 3 435 / **20 %** / 15 512 | verbe, grammaire, langue, corpus, nom, dictionnaire, français, expression, mot, acquisition |
| 14 | Littératures  **25** / 5 142 / **12 %** / 14 278 | roman, poétique, littérature, poésie, fiction, théâtre, lettre, écriture, voix, voyage |
| 15 | Mathématiques  **21** / 888 / **15 %** / 2 745 | cohomologie, package, théorème, régression, algèbre, géométrie, mathématique, assimilation, borne, approximation |
| 16 | Philosophie  **16** / 2 800 / **9 %** / 7 856 | philosophie, épistémologie, critique, idée, éthique, science, réception, concept, logique, vie |
| 17 | Physique  **47** / 3 603 / **48 %** / 30 667 | antenne, spectre, commande, réalisation, couplage, simulation, propriété, détermination, calcul, mesure |
| 18 | Planète et Univers  **13** / 1 245 / **6 %** / 3 675 | géologie, bassin, enregistrement, faune, gisement, légende, datation, variabilité, quantification, fonctionnement |
| 19 | Psychologie  **8** / 943 / **4 %** / 2 663 | autisme, psychanalyse, psychologie, clinique, croyance, différence, enfant, intervention |
| 20 | Science politiques  **23** / 2 520 / **12 %** / 9 864 | parti, élection, démocratie, mobilisation, **Europe**, État, sociologie, justice, politique, acteur |
| 21 | Sciences cognitives  **4** / 1 164 / **2 %** / 3 141 | catégorisation, psychologie, trouble, acquisition |
| 22 | Sciences de l’environnement  **14** / 1 983 / **10 %** / 7 484 | brève, bibliographie, agriculture, valorisation, indicateur, évaluation, changement, conséquence, impact, gestion |
| 23 | Sciences de l’information et de la communication  **17** / 2 053 / **9 %** / 7 523 | média, sémiotique, communication, bibliothèque, médiation, intelligence, appropriation, information, norme, dispositif |
| 24 | Sciences du Vivant  **49** / 3 800 / **37 %** / 22 149 | dosage, composition, influence, effet, intérêt, facteur, conséquence, qualité, variation, utilisation |
| 25 | Sociologie  **40** / 5 268 / **24 %** / 32 398 | géographie, sociologie, territoire, migration, ville, mobilité, paysage, espace, travail, dynamique |

Il faut à présent évaluer ces résultats. Le premier contrôle que nous pouvons effectuer, bien que très subjectif et limité, et de parcourir nous-même ces têtes pour voir si certaines semblent ne pas correspondre au domaine associé. Ce premier contrôle montre que le filtre semble fonctionner : les mots semblent effectivement soit des objets d’études des domaines, comme *céramique* et *nécropole* pour l’archéologie, soit des objets supports de l’activité scientifique comme *étude* ou *approche.*

Une méthode d’évaluation des résultats coûteuse en temps aurait été de soumettre l’ensemble des têtes du corpus à un panel de spécialistes de chaque domaine qui auraient ensuite catégorisé chaque tête comme étant propre à leur domaine, soit de façon binaire, soit sur une échelle. La difficulté matérielle de la tâche et sa part de subjectivité, que l’accord inter-annotateur peut néanmoins corriger, ne nous a pas fait considérer cette option, nous privant des mesures de bruit et de rappel.

On peut essayer de mesurer tout d’abord sa sélectivité : le pourcentage de têtes retenues pour chaque domaine, d’abord pour la condition sur le seuil de fréquence puis sur la condition sur le seuil de distribution. On peut calculer le taux de sélection de têtes, en parlant en lemmes et non en occurrences, pour chaque domaine. Pour la première condition, on a un taux de sélection qui va de près de 0,3 % à 29 %, pour une moyenne de 4 % et une médiane de 2 %, un écart-type de 0,057. Nous observons donc une très grande disparité. L’économie et finance quantitative a un taux de sélection de 29 %, suivi de la géographie avec 10 %, puis la chimie, les mathématiques et la psychologie à 7 %. Cela vient du rapport entre le nombre de lemmes de têtes différents par rapport au nombre total de têtes dans ce domaine. Certains domaines utilisent une grande variété de lemmes pour leurs têtes, dans d’autres les occurrences de têtes sont concentrées sur un plus petit nombre de lemmes. L’économie et finance quantitative utilise 273 lemmes de têtes différents pour 489 occurrences de têtes, soit un ratio de 56 %, la géographie en utilise 604 pour 1 191 occurrences, soit un ratio de 51 %. À l’opposé, la physique utilise 3 584 lemmes pour 30 667 occurrences de têtes en tout, soit un ratio de 12 %.

Pour la seconde condition sur DISTDOMAINE(OCCTÊTE), on calcule le taux de sélection par rapport à l’ensemble retourné par la première sélection. L’étendue est encore plus grande : de 5 % à 98 %. La moyenne est de 51 % et la médiane de 47 % néanmoins. L’écart-type lui confirme bien cette dispersion, il est de 0,31. Ainsi le domaine gestion et management retient 50 des lemmes de têtes sur les 51 retenus par la précédente condition. À l’inverse, l’économie et finance quantitative et la géographie n’en retiennent que 5 %. Cela s’explique notamment par le très faible nombre de titres, 860 pour la géographie, 346 pour l’économie et finance quantitative, et donc d’occurrences de têtes dans les deux domaines.

Pour l’ensemble des conditions, on oscille entre 2,41 % têtes retenues pour la chimie et 2,36 % pour les mathématiques, contre 0,27 % pour l’anthropologie et l’histoire. Notre filtre est donc très sélectif, et semble choisir des têtes adéquates. Mais on peut se poser la question du partage des têtes spécifiques entre les domaines.

Un autre contrôle possible est en effet de mesurer les collisions entre les domaines : une même tête, peut-elle se retrouver spécifique à plusieurs domaines ? Nous calculons donc pour les 355 têtes le nombre de domaines spécifiques auxquels elles sont associées.

|  |  |
| --- | --- |
| **Nombre de domaines associés** | **Nombre de têtes** |
| 6 | 4 1 % |
| 5 | 11 3 % |
| 4 | 19 5 % |
| 3 | 37 10 % |
| 2 | 55 15 % |
| 1 | 230 64 % |

Nous calculons pour cette série le maximum, 6 pour *état, analyse, évaluation* et *outil*, la moyenne, 1,70, la médiane 1,00 et l’écart-type 1,15 de la série obtenue ainsi. Nous constatons que 64 % des têtes sélectionnées ne sont associées qu’à un seul domaine, et 79 % à un ou deux domaines, ce qui est un bon résultat. Nos têtes spécifiques sont donc très peu partagées entre plusieurs domaines, ce qui amoindrirait leur spécificité. Ce partage n’est pas forcément antithétique de la spécificité. Les têtes étant spécifiques à trois domaines sont par exemple *ville,* partagée par les domaines histoire, architecture et sociologie ce qui nous apparaît comme logique, ou encore *architecture* entre architecture, informatique et art et histoire de l’art. Cela nous amène à une première limite de notre approche : la polysémie de certains termes fait que ce lemme *architecture* qui est une tête commune à ces trois domaines ne fait pas référence à la même. L’*architecture* en informatique peut désigner l’architecture des réseaux, des systèmes, des machines, des processeurs, c’est-à-dire leur agencement en vue d’accomplir leurs buts. Cette polysémie qui se cache sous un lemme unique plaide pour une plus grande spécificité réelle que celle déterminée par notre algorithme.

Une seconde limite qui découle de la première est qu’un sens peut être très spécifique à un domaine et un autre nom : synthèse est ainsi partagée entre physique, informatique et chimie. En ce qui concerne la chimie, *synthèse* désigne la création de façon artificielle d’un composé chimique. Mais *synthèse* désigne aussi un support du travail scientifique très générique : il est difficile de savoir à quel sens rattacher les emplois de ce lemme dans les deux autres domaines. Cela concerne surtout les d’objets supports de l’activité scientifique qui peuvent être partagées, ce qui se retrouvent néanmoins dans les nombreuses têtes spécifiques à plusieurs domaines. La tête *analyse* est ainsi présente dans tous les 25 domaines et apparaît comme une tête spécifique pour six domaines : gestion et management, physique, sciences de l’environnement, sciences du vivant, informatique, éducation et sociologie.

Cette remarque nous nous amène à aborder l’autre facette remarquable des têtes de segments : les têtes transdisciplinaires. Obtenir cette liste permettra également de contrôler encore notre liste de têtes spécifiques en croisant les deux.

### II.1.2 Les têtes de segments transdisciplinaires

#### A. Définitions théorique et opératoire

Pour être véritablement transdisciplinaire, une tête ne doit pas seulement se retrouver dans de nombreux domaines. Elle doit se retrouver *fréquemment* dans de nombreux domaines. Nous avons calculé, pour chaque tête et pour chaque domaine, la fréquence des occurrences la tête par rapport au nombre total d’occurrences de têtes dans ce domaine : **FREQ(OCCTÊTE)**. Nous avons donc pour chaque tête sa série de fréquences pour chaque domaine, **FREQ(OCCTÊTE)**.

Pour trouver les têtes transdisciplinaires, nous nous méfions de la moyenne des fréquences de la tête dans les différents domaines par rapport au nombre total de têtes dans ce domaine : **MOYENNE(** **FREQ(OCCTÊTE)** **)**. Une moyenne peut en effet cacher des situations très disparates. Nous regardons donc la médiane de la série **FREQ(OCCTÊTE)** : plus elle sera élevée, plus la tête sera présente fréquemment dans de nombreux domaines. Nous établissons un seuil arbitraire de 0,001 (0,1 %), que nous nommons **seuil de médiane**, au-dessus duquel nous sélectionnons nos têtes transdisciplinaires.

**Filtre de sélection des têtes transdisciplinaires**

**pour chaque** TÊTE **fait  
 si** MEDIANE(FREQ(OCCTÊTE)) > seuil de médiane **alors  
 sélectionne** TÊTE **fin si  
fin pour**

#### B. Résultats et évaluations du résultat

Sur les 123 227 lemmes de têtes de notre corpus de travail, cela en sélectionne 94 soit 0,08 %. Elles ont en tout 94 738 occurrences, soit près de 27 % des 354 168 occurrences de têtes que comptent notre corpus. Les occurrences de ce très petit nombre de têtes transdisciplinaires concentrent plus d’un quart de toutes les têtes.

Les 20 premières têtes des 94 classés par la médiane de la série **FREQ(OCCTÊTE)** sont : *étude, analyse, cas, approche, exemple, enjeu, évolution, apport, rôle, modèle, réflexion, évaluation, outil, question, représentation, application, construction, introduction, histoire* et *développement*. La liste complète est fournie en annexe. Aucun nom propre ne figure dans cette liste ce qui est logique, il s’agit de noms communs abstraits.

Le premier contrôle possible pour tester la validité de notre filtre est de compter les domaines où ces têtes sont présentes. Tutin (2008) fixe la présence d’une forme dans 15 domaines comme marque de sa transdisciplinarité. 15 domaines représentent 60 % des 25 domaines retenus pour nos calculs sur les 27 de notre corpus. Nos 94 têtes transdisciplinaires sont au minimum présentes dans 20 domaines, soit 80 % des 25 domaines. 35 têtes transdisciplinaires sont présentes dans les 25 domaines. Le nombre moyen de domaine où les 94 têtes sont présentes est 23,95 ce qui est extrêmement élevé sachant que le minimum est 20.

Un second contrôle est de le confronter à la liste des noms du lexique transdisciplinaire des écrits scientifiques (LTES) établie par Tutin (2007, 2008). Sur les 94 têtes transdisciplinaires, 74 sont présentes dans le LTES soit 79 %. Les 20 têtes qui ne figurent pas dans le LTES sont : *enjeu, histoire, dynamique, regard, impact, retour, essai, politique, enseignement, note, formation, science, remarque, émergence, point, conception, méthodologie, discours, défi, jeu*. Il nous semble paradoxal que certains lemmes ne figurent pas dans le LTES, surtout ceux sémantiquement liés directement à la science comme *méthodologie* ou *science*. Les autres peuvent avoir été considéré comme trop générique : il en effet difficile de délimiter ce qui est propre à la science, le lexique transdisciplinaire des écrits scientifiques étant considéré comme un sous-ensemble d’un lexique abstrait général (Tutin, 2007).

Un troisième contrôle est de mesurer le croisement entre la liste des têtes spécifiques et les têtes transdisciplinaires. Néanmoins, intuitivement, les deux ensembles ne sont pas forcément disjoints : une tête transdisciplinaire peut être très présente dans une discipline, au point d’en devenir représentative, tout en étant présente dans toutes. Sur nos 94 têtes transdisciplinaires, 89 sont présentes dans la liste des 356 têtes spécifiques, ce qui fait un recouvrement de 95 %. Les 5 têtes qui ne sont pas dans notre précédente classe sont : *an, cadre, défi, enquête* et *perception*. Nous pouvons donc dire que les têtes transdisciplinaires structurent tous les titres, comme l’armature d’un bâtiment dont les domaines seraient les différents étages. Les têtes transdisciplinaires sont tellement présentes qu’elles apparaissent comme spécifique à certains domaines, elles écrasent les spécificités disciplinaires sous leur nombre. Seul le recul apporté par l’étude global de tous les domaines permet de saisir l’image d’ensemble, le fait que l’armature soit bien commune à tous les titres.

#### C. Remarques sur les sous-corpus

Nous avons ensuite étudié les têtes transdisciplinaires sur trois sous-ensembles de notre corpus de travail, les titres monosegmentaux, les premiers segments des titres bisegmentaux, puis leurs seconds segments. Nous traitons les segments des titres bisegmentaux séparément pour essayer de déterminer d’éventuelles différences entre les deux.

Pour les titres monosegmentaux, les têtes transdisciplinaires relevées sont au nombre de 81. Six seulement d’entre elles n’apparaissent pas dans les 94 têtes transdisciplinaires relevés sur tout le corpus. Les six têtes sont : *contrôle, fonction, notion, temps, transformation* et *valeur*. Pour le premier segment des titres bisegmentaux, nous relevons 63 têtes transdisciplinaires. Cinq têtes n’apparaissent pas dans les 94 précédemment relevées : *compte, contribution, culture, économie* et *identité*. Dans le second segment, nous relevons 99 têtes transdisciplinaires et 19 têtes n’apparaissent pas dans les 94 têtes transdisciplinaires relevés sur tout le corpus : *condition, contexte, définition, démarche, donnée, illustration, leçon, limite, mode, mythe, paradoxe, parcours, piste, problématique, réalité, revue, source, synthèse* et *voie*.

Un fait remarquable du sous-corpus de travail des seconds segments de titres bisegmentaux, c’est que certaines têtes transdisciplinaires sont surreprésentées spécifiquement dans ce corpus. Les occurrences des têtes *cas, exemple, étude, application* et *approche* représentent respectivement 4 %, 3 % et 2 % pour les trois dernières des 95 282 occurrences de têtes de ce sous-corpus. Cette très forte présence ne se rencontre pas dans l’ensemble du corpus et le corpus des premiers segments des titres bisegmentaux. Les occurrences de la tête *étude* du corpus de travail ne représente que 2 % du total des occurrences de têtes, celles des têtes analyse et étude près de 1 % du corpus des premiers segments des titres bisegmentaux. Uniquement voit-on dans le sous-corpus des titres monosegmentaux poindre *étude* à 3 %. Il y a donc une concentration remarquable sur un petit nombre de têtes dans le sous-corpus des seconds segments de titres bisegmentaux.

On peut également étudier les titres bisegmentaux en prenant les deux têtes ensembles, formant ainsi des couples ordonnés de la forme (tête premier segment, tête second segment). Seuls cinq couples ont une médiane différente de 0 : *(de, exemple), (rôle, cas), (approche, cas), (apport, exemple) et (effet, cas)*. L’apparition de la préposition *de* s’explique car nous exigeons qu’une des têtes soient un nom, l’autre peut être un verbe ou une préposition. La préposition *de* étant la plus fréquente, il est logique qu’elle apparaisse dans les couples les plus fréquents. Cette préposition est utilisée dans des structures de la forme *de … vers …*, *de … à …* étudiées par Tanguy et Rebeyrolle (à paraître), mais comme il ne s’agit pas d’un nom nous l’écartons ici. Un autre aspect remarquable est l’existence de 409 titres dont le premier et le second segment ont le même lemme pour tête, pour achever un effet stylistique de répétition et introduire une comparaison ou un questionnement :

(11) Titre n°271743 : La **crise** ? Quelle **crise** ?

(12) Titre n°1735179 : **Crise** du logement ? Quelle **crise** ?

(13) Titre n°1522267 : **Ville** de jour. **Ville** de nuit

(14) Titre n°183060 : **Linux** embarqué. **Linux** Temps Réel

(15) Titre n°594563 : **Feu** l'arrêt Mercier ! **Feu** l'arrêt Mercier ?

(16) Titre n°740925 : **Corps** dansant. **Corps** glorieux

#### Conclusion de la recherche de têtes spécifiques et transdisciplinaires

Après avoir regardé le corpus de travail dans son ensemble et séparément en sous-corpus, nous avons fait émerger d’un côté des têtes spécifiques à des disciplines et de l’autre des têtes transdisciplinaires. Nous avons sélectionné 356 têtes spécifiques pour les 25 domaines différents. Nous avons néanmoins remarqué un fort recouvrement avec la seconde classe que nous avons fait émerger : celle des têtes transdisciplinaires. En effet, 95 % des têtes spécifiques sont aussi transdisciplinaires selon nos filtres.

Nous avons dans cette partie identifié un petit nombre de têtes transdisciplinaires, 121 en tout si on reprend tous les lemmes identifiés dans les différents sous-corpus, 94 dans le corpus de travail général, très fréquentes et donc utilisées dans de nombreux titres de notre corpus de travail et, pour à 86 % pour les 121 têtes et à 79 % pour les 94, déjà relevé dans le lexique transdisciplinaire des écrits scientifiques de Tutin (2008). L’étude du second segment des titres bisegmentaux a mis en avant deux têtes transdisciplinaires qui le caractérisent tout particulièrement, *cas* et *exemple*.

Les têtes transdisciplinaires sont caractérisées par une haute fréquence en tant que têtes et en tant que simples noms et un haut degré d’abstraction. Du fait de leur caractère abstrait et de leur transdisciplinarité, on peut s’interroger sur l’importance de leur contenu sémantique. Que référence-t-on exactement lorsque l’on parle d’une *étude* ou d’un *cas*, d’un *outil* ou d’une *contribution* ? Nous devons ici faire un détour pour présenter un emploi nominal particulier, celui de nom général sous-spécifié, et en quoi cet emploi peut se rapprocher de notre classe de têtes transdisciplinaires.

## II.2 Les noms généraux sous-spécifiés

### II.2.1 Principes théoriques

De nombreux travaux se sont penchés sur les noms généraux sous-spécifiés (NGSS) en anglais et plus tardivement en français. Si les travaux s’accordent pour définir les NGSS comme un emploi particulier, les définitions théoriques et opératoires de cet emploi sont sujettes à débat, ainsi que la liste des noms pouvant être employé de la sorte, comme le reflète le foisonnement terminologique pour désigner cet emploi : *signalling nouns* (Flowerdew 2003, 2006 ; Flowerdew et Forest, 2015), *type 3 vocabulary* (Winter, 1977), *metadiscursive nouns* ou *anaphoric nouns* (Francis, 1986), *enumerables* et *advance labels* (Tadros, 1994), *carrier nouns* (Ivanic, 1991), *advance labels* et *retrospective labels* (Francis, 1994), *unspecific nouns* ou *metalanguage nouns* (Winter, 1992), *shell nouns* (Hunston et Francis, 1999 ; Schmid, 2000, 2018), *noms sous-spécifiés* (Legallois, 2008) et *noms porteurs* (Huygue, 2018).

Comme définition théorique, nous nous proposons de reprendre celle de Flowerdew (2006) pour sa concision et sa clarté (nous traduisons) : *« noms abstraits dont le sens complet peut seulement être spécifié en référence à son contexte »*. Un exemple d’emploi sous-spécifié pour le lemme *défi* est le suivant : “Pour les Américains, le **défi** est de *marcher à nouveau sur la Lune mais cette fois-ci pour la conquérir*”. Le sens complet de *défi* ne peut être appréhendé qu’en faisant référence à au contexte, ici “*marcher à nouveau sur la Lune mais cette fois-ci pour la conquérir”*.

Pour compléter notre définition théorique, on rappellera également les trois fonctions clés de l’emploi sous-spécifié selon Schmid (2000) :

* Capacité de référence quasi-pronomominale
* Création de concepts temporaires
* Catégorisations de concepts par le locuteur

Devant le faible nombre de couples sélectionnés par notre filtre, qui n’est pas forcément adapté aux couples, on peut regarder la moyenne et sélectionner les meilleurs couples : *(modèle, application), (analyse, application), (impact, cas),* *(méthode, application), (modélisation, application), (étude, application), (utilisation, application), (évolution, analyse)*.

Ce qui apparaît clairement à la lecture de ces couples transdisciplinaires restants est un lien sémantique entre les deux têtes. Pour s’en rendre mieux compte, voici quelques exemples de titres avec les têtes mises en gras :

(11) Titre n°1123145 : Le **rôle** des territoires de projets infra-régionaux sur l’agriculture et l’alimentation. Le **cas** du développement de circuits de proximité

(12) Titre n°1634661 : Une **approche** cognitive de la performance d’un réseau : le **cas** d’un Plan Local d’Insertion par l’Economique

(13) Titre n°**467868** : **Apports** de l'imagerie satellitale thermique à la connaissance de la dynamique océanique : **exemple** de la Méditerranée nord-occidentale

(14) Titre n°**1190686** : **Effet** de la température sur le devenir des pesticides dans les sols et conséquences pour l’évaluation des risques environnementaux : **cas** du glyphosate et des herbicides sélectifs

(15) Titre n°110970 : Un **modèle** générique d'organisation de corpus en ligne : **application** à la FReeBank

(16) Titre n°1516462 : **Analyse** numérique d'une séquence d'images ultra-rapides: **application** à la fissuration dynamique de structures impactées

(17) Titre n°1201029 : **Impact** des variables et pratiques agronomiques sur la réduction des dommages: le **cas** de la pomme de terre au Québec

(18) Titre n°615757 : Une **méthode** générale pour construire automatiquement des atlas anatomiques morphométriques : **application** à un atlas du crâne

(19) Titre n°362242 : **Modélisation** des procédés de nanofiltration avec le logiciel NanoFlux : **applications** nucléaires

(20) Titre n°158164 : **Etude** expérimentale de l'inversion thermique en radiométrie micro-onde par corrélation : **applications** médicales

(21) Titre n°353524 : L'**utilisation** de caractères morphologiques dans la diagnose sexuelle des mandibules d'enfants : **application** à la collection ostéologique de Turin (Italie)

(22) Titre n°39219 : **Utilisation** d'un carbone activé comme support de catalyseur. **Application** à l'hydrogénation du glucose

(23) Titre n°522578 : **Evolution** des règles d'utilisation du soja transgénique au Brésil : une **analyse** par une approche systémique de la gouvernance

(24) Titre n°589931 : **Evolution** des températures à Nantes entre 1946 et 2000. **Analyses** statistiques des ruptures de tendance et interprétations

Les couples d’exemples 21 et 22 et 23 et 24 montrent la supériorité de l’approche mise en place pour ce travail par rapport à notre mémoire de M1 : malgré la différence de segmentation, faible pour 21 et 23 avec le double point et forte pour 22 et 24 avec un point, on retrouve une même structure transsegmentale détectée par nos recherches.

Une question qui apparaît est la détermination des têtes où trois cas sont possibles : pas de déterminant, comme dans la première tête de l’exemple 20, un déterminant défini comme dans la seconde tête de l’exemple 11 ou un déterminant indéfini comme pour la seconde tête de l’exemple 23. Le choix du déterminant est crucial pour l’interprétation sémantique du titre, comme le montrent les deux exemples suivants :

(25) Titre n°828580 : **Madagascar** : un **cas** d'école pour le Développement durable

(26) Titre n°62678 : **Exposer** des objets sonores : le **cas** des chansons de Brassens

Dans l’exemple 25, *un cas* indique que Madagascar est pris comme exemple pour le concept de *développement durable*. Dans l’exemple 26, *le cas* indique que ce sont les chansons de Brassens qui sont pris comme exemple du concept *exposer des objets sonores*. Cette interprétation sémantique opposée selon le déterminant amène à s’interroger sur les déterminants des têtes de nos couples. Nous prenons ici les quatre couples ayant uniquement des noms comme têtes *(rôle, cas), (approche, cas), (apport, exemple) et (effet, cas)* :

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Tête n°1** | **Sans** | **Déf.** | **Indéf.** | **Tête n°2** | **Sans** | **Déf.** | **Indéf.** |
| rôle | 11 | 42 | 0 | cas | 5 | 48 | 0 |
| approche | 18 | 4 | 19 | cas | 9 | 31 | 0 |
| apport | 0 | 17 | 0 | exemple | 10 | 21 | 1 |
| Effet | 18 | 18 | 0 | cas | 10 | 26 | 0 |

La distribution entre les trois possibilités, aucun déterminant, déterminant défini et déterminant indéfini montre qu’il existe des contraintes sur ce choix dans les couples de têtes. Ainsi, le couple *(rôle, cas)* sera à 70 % *(le rôle, le cas)* et on ne trouvera jamais *un cas* dans un couple avec *rôle* en première tête. Pour les autres couples, la majorité des occurrences n’emploient pas de déterminant sauf dans le couple *(méthode, application)* qui a près de 50 % utilise une déterminant indéfini pour *méthode.*

Ce lien sémantique nous amène à aborder ce que l’on trouve dans les titres sur ce plan. D’après Anthony (2001), Cheng et al. (2012) et Grant (2013), nous retrouvons neuf grandes catégories d’informations sémantiques au sein des titres : sujet, focalisation du sujet, méthode employée, description, nom d’un procécédé, d’une technique ou d’un outil, source des données utilisées, métaphore entre deux objets, question et les résultats de l’étude. Ces mêmes travaux énumèrent des couples de catégories dont on peut relier deux d’entre elles, présentes dans Anthony (2001) et Cheng et al. (2012), à nos couples de têtes en observant manuellement les différents titres, avec une variation d’ordre pour le couple formé par les catégories sujet et méthode :

* Sujet - focalisation : *(rôle, cas), (modélisation, influence), (intégration, cas), (impact, cas), (de, exemple), (de, cas), (approche, cas), (étude, application), (étude, cas), (effet, cas)*
* Méthode - sujet : *(méthode, application), (analyse, application), (modélisation, application), (modèle, application), (utilisation, application)*
* Sujet - méthode : *(évolution, étude), (rôle, approche)*

TODO

Les titres, pris isolément, sont des microdiscours où la capacité de référence quasi-pronominale de l’emploi sous-spécifié de noms n’est pas visible, l’espace étant trop court pour une reprise en anaphore ou cataphore comme dans l’exemple : « *Le problème des échanges commerciaux Chine-États-Unis. Un problème vital »*.

Or, cette capacité de référence est une des trois fonctions clés de l’emploi sous-spécifié de noms selon Schmid (2000), avec la création de concepts temporaires et la catégorisation de ces concepts par le locuteur, qui elles sont bien présentes. Qualifier les échanges commerciaux dans l’exemple précédent de “problème” ou de “question” est une catégorisation qu’impose le locuteur sur la perception du contenu spécifiant. Un concept temporaire est bien formé mentalement dans la tête du lecteur, celui des échanges commerciaux entre la Chine et les États-Unis.

Enfin, on aborde le fait qu’il s’insère dans une construction spécificationnelle (CS).

### II.2.2 Présence des constructions spécificationnelles dans notre corpus

TODO

Établir un lexique des têtes les plus fréquentes sur les 15 692 lemmes recensés, proposer quelques remarques générales dessus notamment l’appartenance au vocabulaire académique transdisciplinaire (Hatier, 2016 ; Hatier et al., 2016) et le fait que cela soit des noms généraux définis par Halliday et Hasan (1976) et étudiés notamment par Adler et Moline (2018). La question qui est en suspens est : sont-ils des NSS ?

Nous recensons ici les différentes *constructions spécificationnelles* (Legallois, 2008) traditionnelles de la littérature sur les NSS, qui sont autant de définitions opératoires des NSS (Schmid 2000), en essayant de les chercher dans notre corpus. Le lien avec les *grammaires de construction* sera également évoqué (François et Legallois, 2006). Nous commençons par les deux plus étudiées (Legallois, 2008 ; Schmid, 2000) :

* **NSS** + [verbe être] + *proposition subordonnée complétive attribut du sujet* : “le **problème** est *que l’homme souhaite toujours plus*”,
* **NSS** + [verbe être] + **de** + *proposition subordonnée infinitive* : “le **problème** est **de** *délimiter nos souhaits*”.

Nakamura (2017) ajoute également les trois constructions spécificationnelles suivantes :

* **NSS** + [verbe être] + *syntagme nominal* : “Notre **objectif** majeur est *la rédaction d’une proposition de loi*.”
* Nom + [verbe avoir] + pour + **NSS** + **de** + *proposition subordonnée infinitive* : “Cet homme avait pour **ambition** *de devenir président*”.
* **NSS** + de + *syntagme verbal à l’infinitif* : “L’**ambition** *de devenir président*”. Pour le citer “il s’agit de la formation d’un syntagme nominal complexe, qui comporte à la fois la partie sous-spécifiée et la partie spécifiante”.

TODO

Schmid (2018) indique que son étude n’a pris que les deux premières définitions pour des raisons techniques, mais il atteste dès son livre de 2000 l’existence de la troisième CS décrite par Nakamura, que Flowerdew et Forest (2015) évoquent également.

En recherchant dans notre corpus les occurrences de ces constructions spécificationnelles, on montre qu’on ne les trouve pas (TODO : calculer leur faible présence). C’est à partir de la dernière de Nakamura (2017), soutenu par Schmid (2018), que nous avons argument fort pour rapprocher nos noms des NSS : les titres sont des syntagmes nominaux complexes, comme l’est la cinquième construction spécificationnelle. Nous pourrions avoir un emploi de NSS sans verbe à l’infinitif dans les titres. Nous allons donc essayer d’établir des caractéristiques de têtes de segments pour soutenir cet éventuel rapprochement.

## II.3 Caractéristiques retenues de rapprochement

TODO

Nous listons dans cette partie une liste de caractéristiques retenues pour rapprocher les têtes de segments des NSS :

* Fréquence par rapport à l’ensemble des noms.
* S’il s’agit d’un nom abstrait oui ou non.
* Détermination : non définie, définie, pas de détermination.
* Nombre : pluriel ou singulier.
* Complémentation du nom, soit par un syntagme prépositionnel introduit par une autre préposition que *de,* par un groupe introduit par *de* et sans complémentation. Cheng et al. (2012) indique que 90 % des modificateurs des noms sont des groupes prépositionnels, ce qui est une caractéristique de l’écriture académique (Biber et al., 1999 ; Biber et Gray, 2010), et que la majorité de ces groupes utilisent *of* ou *in* comme préposition.
* Transdisciplinarité : Moyenne de la position du lemme dans le classement en fréquence dans les différents domaines. Plus elle sera haute, plus sa transdisciplinarité sera bonne. On fera attention de distinguer sur le sens : sémantique neutre (“modèle”, “analyse”), sémantique en rapport avec un seul domaine (“architecture”), sémantique en rapport avec plusieurs domaines (“histoire”, “ville”), sémantique mixte interprétable comme neutre mais aussi comme en rapport avec un seul domaine (“synthèse”).
* Appartenance à la liste établie par Flowerdew et Forest (2015).
* Appartenance à la liste établie par Scmid (2000).
* Position de la racine dans leur segment. Roze et al. (2014) indique l’existence d’un schéma *Nom sous-spécifié : suite*, ce qui laisse à penser que la position des racines est importante, dans ce juste avant le signe de ponctuation segmentant.
* Position de leur segment dans le titre par rapport aux autres segments.
* Position de la racine dans le titre.

Exemple :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Lemme | Abstrait | Dét. | Compl. | Transdisciplinarité | NSS fréquent ? | Pos. |
| cas | TODO |  |  |  |  |  |
| problème |  |  |  |  |  |  |
| objectif |  |  |  |  |  |  |

## II.4 Points communs et différences avec les noms sous-spécifiés

### II.4.1 Facteurs de rapprochement

TODO

Établir une liste de facteurs de rapprochement avec les noms sous-spécifiés et les justifier.

Citer des têtes dont on est sûr qu’ils sont des NSS.

### II.4.2 Facteurs de différenciation

TODO

Établir une liste de facteurs de différenciation avec les noms sous-spécifiés et les justifier.

Citer des têtes dont on est sûr qu’ils ne sont pas des NSS.

### II.4.3 Règle de rapprochement

TODO

Essayer d’aboutir à une règle pour dire si une tête est un NSS ou non NSS et proposer une liste des têtes selon cette règle.

# III. Schémas récurrents et constructions spécificationnelles

## III.1 Émergences de schémas fréquents récurrents

Nous étudions dans cette partie les schémas, dans lesquels s’insèrent nos têtes de segments, à l’intérieur d’un seul segment. Le but est d’essayer d’étudier plus en profondeur, au-delà de savoir qu’il s’agit d’un syntagme nominal complexe, l’emploi des têtes de segments et l’émergence de schémas fréquents récurrents. Nous commençons par étudier les schémas intrasegmentaires avant d’aborder les schémas s’étendant sur deux segments.

### III.1.1 Schémas intrasegmentaires

TODO

Par exemple, nous voulons étudier ce qu’implique la différence entre les déterminants “le” et “un” comme dans “le cas de” et “un cas de”, notamment sur la position du segment incluant ce schéma dans le titre et la relation sémantique qui s’établit avec le nom qui vient après. Y’a-t-il des préférences marquées pour la forme définie ou la forme indéfinie pour des noms donnés ?

Y’a-t-il des NSS substituables ? Si on prend “un cas de” et un “un problème de”, peut-on faire des rapprochements entre les deux schémas suivant leur position dans le titre ou le nom qui suit ?

Quelle sémantique peut-on associer à ces schémas ?

### III.1.2 Schémas sur deux segments

TODO

Nous étudions dans cette partie les schémas incluant une ou deux racines s’étendant sur plus d’un segment. On s’attardera notamment sur le cas de la complémentation par un segment de la forme “NSS : contenu spécifiant”.

Nous essayerons là aussi d’aborder la sémantique des schémas, en faisant référence à ce que l’on trouve dans un titre (Grant, 2013 ; Paiva, 2012), mais en nous rapprochant également des typologies sémantiques que l’on plaque sur les titres bisegmentaux comme dans la suite de travaux de Swales et Feak (1994), Anthony (2001) et Cheng et al. (2012) qui partagent une orientation commune.

## III.2 Transdisciplinarité des schémas et schémas non centrés sur têtes

### III.2.1 Transdisciplinarité des schémas

TODO

Dans cette partie nous étudions la répartition des schémas selon les disciplines.

### III.2.2 Schémas non centrés sur une tête

TODO

Dans cette partie nous étudions si nous retrouvons nos schémas centrés non pas sur une tête de segment que nous aurions détectée mais sur d’autres noms. Cela pour étudier si les schémas sont propres aux têtes de segment ou non, éventuellement aux têtes de segment NSS.

## III.3 Des schémas de constructions spécificationnelles ?

TODO

De la même manière que nous avons essayé de rapprocher têtes de segments et NSS, nous voulons voir si nous pouvons rapprocher les schémas récurrents dans lesquels les têtes s’insèrent et les constructions spécificationnelles. Nous pensons à un rapprochement plus lâche que celui entre têtes et NSS qui pourraient néanmoins relever des similitudes dans le fonctionnement sémantique dynamique. Nous pensons notamment aux travaux de Nakamura (2017) sur ce point.

Après avoir étudié les schémas récurrents où se trouvent les têtes de segments, notamment leur transdisciplinarité et leur éventuel rapprochement avec les constructions spécificationnelles, nous voulons discuter des limites de notre travail et des perspectives qu’il ouvre.

# IV. Discussion sur nos résultats, limites et perspectives

A quoi peut servir la liste de mots trans ? ⇒ faire émerger des topics en cours

Aurait-on pu utiliser le machine learning ?

## IV.1 Limites de notre travail

TODO

Citer des têtes dont ne sait pas s’ils sont bien des NSS ou pas.

Les titres complexes avec plus de deux segments.

Limites techniques de l’outil : Talismane, à la manière d’une lunette astronomique, a rendu possible la perception de phénomènes dans notre corpus qui, en retour, demandes des hypothèses explicatives. Mais l’outil est faillible et nous devons en être conscient pour ne pas lui être inféodé (cas du mauvais élément régisseur dans une relation de dépendance par exemple).

Calcul des seuils en fonction des collisions : dynamique jusqu’à trouver les bons

Utiliser les mots pour catégoriser : mais il y en a si peu… Couverture des titres ?

## IV.2 Perspectives

TODO

Sémantique distributive pour étudier les compléments nominaux des NSS.

# Conclusion

TODO

Annexes

# A1. Références bibliographiques

Adler, S. et Moline, E. (2018). Les noms généraux: présentation. *Langue française, 2018(2)*, 5-18.

Aleixandre-Benavent, R., Montalt-Resurecció, V. et Valderrama-Zurián, J. (2014). A descriptive study of inaccuracy in article titles on bibliometrics published in biomedical journals. *Scientometrics, 101(1)*, 781-791.

Anthony, L. (2001). Characteristic features of research article titles in computer science. *IEEE   
Transactions on Professional Communication, 44(3)*, 187-194.

Ball, R. (2009). Scholarly communication in transition: The use of question marks in the titles of scientific articles in medicine, life sciences and physics 1966–2005. *Scientometrics, 79(3)*, 667-679.

Baethge, C. (2008). Publish together or perish: the increasing number of authors per article in academic journals is the consequence of a changing scientific culture. *Deutsches Arzteblatt international, 105(20)*, 380-383.

Cheng, S. W., Kuo, C. W. et Kuo, C. H. (2012). Research article titles in applied linguistics. *Journal of Academic Language and Learning, 6(1)*, A1-A14.

Cori, M. et David, S. (2008). Les corpus fondent-ils une nouvelle linguistique ? *Langages, 171*, 111-129.

Diers, D. et Downs, F. S. (1994). Colonizing: a measurement of the development of a profession. *Nursing research, 43(5)*, 316.

Dillon, J. (1981). The emergence of the colon: an empirical correlate of scholarship. *American Psychologist, 36*, 879-884.

Dillon, J. T. (1982). In Pursuit of the Colon, A Century of Scholarly Progress: 1880–1980. *The Journal of Higher Education, 53(1)*.

Flowerdew, J. (2003). Signalling nouns in discourse. *English for specific purposes, 22(4)*, 329-346.

Flowerdew, J. (2006). Use of signalling nouns in a learner corpus. *International Journal of Corpus Linguistics, 11(3)*, 345-362.

Flowerdew, J. & Forest, R. W. (2015). *Signalling nouns in English*. Cambridge University Press.

Francis, G. (1986). *Anaphoric nouns*. English Language Research, Department of English, University of Birmingham.

Francis, G. (1994). Labelling discourse: an aspect of nominal-group lexical cohesion. In Coulthard, M. ed, (1994), *Advances in written text analysis*, London: Routledge, 83-101.

François, J. et Legallois, D. (2006). Autour des grammaires de constructions et de patterns. *Cahiers du CRISCO.* Université de Caen.

Goodman, R. A., Thacker, S. B. et Siegel, P. Z. (2001). What’s in a title? A descriptive study of article titles in peer-reviewed medical journals. *Science, 24(3)*, 75-78.

Grant, M. J. (2013). What makes a good title? *Health Information & Libraries Journal, 30(4)*, 259-260.

Gustavii, B. (2017). *How to write and illustrate a scientific paper*. Cambridge University Press.

Haggan, M. (2004). Research paper titles in literature, linguistics and science: dimensions of attraction. *Journal of Pragmatics, 36(2)*, 293-317.

Hallliday, M. A. K. et Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hartley, J. (2005). To attract or to inform: What are titles for? *Journal of technical writing and communication, 35(2)*, 203-213.

Hatier, S. (2016). Identification et analyse linguistique du lexique scientifique transdisciplinaire. Approche fouillée sur corpus d’article de recherche en SHS, Thèse de doctorat, Université Grenoble Alpes, 2016.

Hatier, S., Augustyn, M., Tran, T. T. H., Yan, R., Tutin, A. & Jacques, M. P. (2016). French cross-disciplinary scientific lexicon: extraction and linguistic analysis. In *Proceedings of Euralex*, 355-366.

Ho-Dac, L.-M., Jacques, M.-P. & Rebeyrolle, J. (2004). Sur la fonction discursive des titres. Dans S. Porhiel et D. Klingler (éds). *L’unité texte*, Pleyben, Perspectives, 125-152.

Hunston, S. & Francis, G. (1999). *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: Benjamins (Studies in Corpus Linguistics 4).

Huyghe, R. (2018). Généralité sémantique et portage propositionnel: le cas de fait. *Langue française, 2018(2)*, 35-50.

Ivanic, R. (1991). Nouns in search of a context: A study of nouns with both open- and closed-system characteristics. *International Review of Applied Linguistics in Language Teaching, 2*, 93-114.

Jacques, T. S. et Sebire, N. J. (2010). The impact of article titles on citation hits: an analysis of general and specialist medical journals. *Journal of the Royal Society of Medicine Short Reports, 1(1)*, 1-5.

Jamali, H. R. et Nikzad, M. (2011). Article title type and its relation with the number of downloads and citations. *Scientometrics, 88(2)*, 653-661.

Kutch, T. D. C. (1978). Relation of title length to numbers of authors in journal articles. *Journal of the American Society of Information Science, 19(4)*, 200-202.

Larivière, V., Gingras, Y., Sugimoto, C. R. and Tsou, A. (2015). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology, 66(7)*, 1323-1332.

Leech, G. N. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning, 50(4)*, 675-724.

Legallois, D. (2008). Sur quelques caractéristiques des noms sous-spécifiés. *Scolia, 23*, 109-127.

Mabe, M. A. et Amin, M. (2002). Dr. Jekyll and Dr. Hyde: Author-reader asymmetries in scholarly publishing. *Aslib Proceedings, 54(3)*, 149-157.

Merrill, E., & Knipps, A. (2014). What's in a Title?. *The Journal of Wildlife Management, 78(5)*, 761-762.

Nagano, R. L. (2015). Research article titles and disciplinary conventions: A corpus study of eight disciplines. Journal of Academic Writing, 5(1), 133-144.

Nakamura, T. (2017). Extensions transitives de constructions spécificationnelles. *Langue française, 2017 (2)*, 69-84.

Nivard, J. (2010). Les Archives ouvertes de l’EHESS. Récupéré sur *La Lettre de l'École des hautes études en sciences sociales n°34*:<http://lettre.ehess.fr/index.php?5883>

Paiva, C. E., Lima, J. P. da S. N. et Paiva, B. S. R. (2012). Articles with short titles describing the results are cited more often. *Clinics, 67(5)*, 509-513.

Rebeyrolle, J., Jacques, M. et Péry-Woodley, M. (2009). Titres et intertitres dans l’organisation du discours. *Journal of French Language Studies, 19*, 269-290.

Roze, C., Charnois, T., Legallois, D., Ferrari, S. et Salles, M. (2014). Identification des noms sous-spécifiés, signaux de l’organisation discursive. Dans *Proceedings of TALN 2014, 1,* 377-388.

Sagi, I., & Yechiam, E. (2008). Amusing titles in scientific journals and article citation. *Journal of Information Science, 34(5)*, 680-687.

Salager-Meyer, F. & Alcaraz Ariza, M. Á. (2013). Titles are" serious stuff": a historical study of academic titles. *Jahr, 4(7), 257-271.*

Schmid, H.-J. (2000). *English Abstract Nouns as Conceptual Shells. From Corpus to Cognition*. Berlin: Mouton de Gruyter (Topics in English Linguistics 34).

Schmid, H. J. (2018). Shell nouns in English-a personal roundup. *Caplletra. Revista Internacional de Filologia*, (64), 109-128.

Schwischay, B. (2001). Notes d’exposés sur deux modèles de description syntaxique [Document PDF]. Repéré à <http://www.home.uni-osnabrueck.de/bschwisc/archives/deuxmodeles.pdf>

Soler, V. (2007). Writing titles in science: An exploratory study. *English for Specific Purposes, 26*, 90–102.

Soler, V. (2011). Comparative and contrastive observations on scientific titles written in English and Spanish. *English for Specific Purposes, 30(2)*, 124-137.

Subotic, S. & Mukherjee, B. (2014). Short and amusing: The relationship between title characteristics, downloads, and citations in psychology articles*. Journal of Information Science, 40(1)*, 115-124.

Swales, J. M. et Feak, C. B. (1994). *Academic Writing for Graduate Students*. Ann Arbor: University of Michigan Press.

Tadros, A. (1994). Predictive categories in expository text. In Coulthard, M. ed, (1994), *Advances in written text analysis*, London: Routledge, 83-96.

Tanguy, L., Rebeyrolle, J. (à paraître). Les titres des publications scientifiques en français : fouille de texte pour le réperage de schémas lexico-syntaxiques.

Townsend, M. A. (1983). Titular Colonicity and Scholarship: New Zealand Research and Scholarly Impact. *New Zealand Journal of Psychology, 12*, 41-43.

Tutin, A. (2007). Autour du lexique et de la phraséologie des écrits scientifiques. *Revue Française de Linguistique Appliquée, 12(2)*, 5-14.

Tutin, A. (2008). Sémantique lexicale et corpus : l’étude du lexique transdisciplinaire des écrits scientifiques. *Lublin studies in modern languages and litterature, 32,* 242-260.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Toulouse: Doctoral dissertation, Université de Toulouse II-Le Mirail.

Urieli, A. et Tanguy, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. *Actes de TALN*, Sables D'Olonne.

Wang, Y. et Bai, Y. (2007). A corpus-based syntactic study of medical research article titles. *System,* *35(3)*, 388-399.

Winter, E. O. (1977). A clause-relational approach to English texts: a study of some predictive lexical items in written discourse. *Instructional science*, *6*(1), 1-92.

Winter, E. O. (1992). The notion of unspecific versus specific as one way of analysing the information of a fund-raising letter. *Discourse description: Diverse linguistic analyses of a fund-raising text*, 131-170.

Yitzhaki, M. (1994). Relation of title length of journal articles to number of authors. *Scientometrics, 30(1)*, 321-332.

Yitzhaki, M. (2002). Relation of the title length of a journal article to the length of the article. *Scientometrics, 54(3)*, 435-447.

# A2. Liste des têtes

## A.2.1 Liste des têtes spécifiques aux domaines

for domain 0.math we select 206 heads / 888 ( 23.2 )

for domain 0.math we kept 38 heads / 888 ( 4.28 )

lemma POS F / dom F / o occ [ / dom / occ

estimation NC 0.037158 0.176 102 [ / 2745 / 579 ] 4

théorème NC 0.009107 0.568 25 [ / 2745 / 44 ] 1 old

géométrie NC 0.008015 0.262 22 [ / 2745 / 84 ] 1

mathématique NC 0.007286 0.267 20 [ / 2745 / 75 ] 3

approximation NC 0.005100 0.212 14 [ / 2745 / 66 ] 2

régression NC 0.005100 0.378 14 [ / 2745 / 37 ] 1

package NC 0.004372 0.706 12 [ / 2745 / 17 ] 1 old

cohomologie NC 0.004007 1.000 11 [ / 2745 / 11 ] 1 old

statistique NC 0.004007 0.159 11 [ / 2745 / 69 ] 1

assimilation NC 0.004007 0.244 11 [ / 2745 / 45 ] 1

existence NC 0.003279 0.200 9 [ / 2745 / 45 ] 1

généralisation NC 0.003279 0.155 9 [ / 2745 / 58 ] 1

algèbre NC 0.003279 0.321 9 [ / 2745 / 28 ] 1

borne NC 0.003279 0.225 9 [ / 2745 / 40 ] 2

nombre NC 0.002914 0.222 8 [ / 2745 / 36 ] 1

inférence NC 0.002914 0.190 8 [ / 2745 / 42 ] 1

mélange NC 0.002914 0.151 8 [ / 2745 / 53 ] 1

somme NC 0.002186 0.261 6 [ / 2745 / 23 ] 1

estimateur NC 0.001821 0.312 5 [ / 2745 / 16 ] 1

clustering NC 0.001821 0.357 5 [ / 2745 / 14 ] 1

feuilletage NC 0.001821 0.714 5 [ / 2745 / 7 ] 1 old

lemme NC 0.001821 0.625 5 [ / 2745 / 8 ] 1 old

compactification NC 0.001457 1.000 4 [ / 2745 / 4 ] 1 old

interpolation NC 0.001457 0.222 4 [ / 2745 / 18 ] 1

dérivation NC 0.001457 0.154 4 [ / 2745 / 26 ] 1

minoration NC 0.001457 0.800 4 [ / 2745 / 5 ] 1 old

Henri Poincaré NPP 0.001457 0.200 4 [ / 2745 / 20 ] 2

Mixmod NPP 0.001457 1.000 4 [ / 2745 / 4 ] 1 old

ondelette NC 0.001457 0.571 4 [ / 2745 / 7 ] 1 old

filtration NC 0.001457 0.250 4 [ / 2745 / 16 ] 2

combinatoire NC 0.001093 0.375 3 [ / 2745 / 8 ] 1

conjecture NC 0.001093 0.214 3 [ / 2745 / 14 ] 1

Leonhard ?Euler NPP 0.001093 1.000 3 [ / 2745 / 3 ] 1 old

intervalle NC 0.001093 0.500 3 [ / 2745 / 6 ] 1 old

imputation NC 0.001093 0.231 3 [ / 2745 / 13 ] 1

consistance NC 0.001093 0.231 3 [ / 2745 / 13 ] 1

récréation NC 0.001093 0.500 3 [ / 2745 / 6 ] 1 old

?Cartomagie NPP 0.001093 1.000 3 [ / 2745 / 3 ] 1 old

for domain 1.shs.infocom we select 199 heads / 2028 ( 9.81 )

for domain 1.shs.infocom we kept 28 heads / 2028 ( 1.38 )

lemma POS F / dom F / o occ [ / dom / occ

communication NC 0.010767 0.260 81 [ / 7523 / 312 ] 2

média NC 0.006912 0.531 52 [ / 7523 / 98 ] 1 old

médiation NC 0.006115 0.232 46 [ / 7523 / 198 ] 1

bibliothèque NC 0.005051 0.235 38 [ / 7523 / 162 ] 2

intelligence NC 0.003855 0.225 29 [ / 7523 / 129 ] 2

sémiotique NC 0.003323 0.403 25 [ / 7523 / 62 ] 1

internet NC 0.002924 0.237 22 [ / 7523 / 93 ] 1

web NC 0.002924 0.415 22 [ / 7523 / 53 ] 1

document NC 0.002659 0.183 20 [ / 7523 / 109 ] 1

presse NC 0.002659 0.253 20 [ / 7523 / 79 ] 1

journaliste NC 0.002659 0.500 20 [ / 7523 / 40 ] 1 old

télévision NC 0.002526 0.487 19 [ / 7523 / 39 ] 1

design NC 0.002526 0.200 19 [ / 7523 / 95 ] 1

journalisme NC 0.002260 0.500 17 [ / 7523 / 34 ] 1 old

tic NC 0.002260 0.279 17 [ / 7523 / 61 ] 1

blog NC 0.002260 0.586 17 [ / 7523 / 29 ] 1 old

médiatisation NC 0.001861 0.326 14 [ / 7523 / 43 ] 1

créativité NC 0.001861 0.233 14 [ / 7523 / 60 ] 1

carnet NC 0.001595 0.333 12 [ / 7523 / 36 ] 1

public NC 0.001329 0.250 10 [ / 7523 / 40 ] 1

open NC 0.001329 0.476 10 [ / 7523 / 21 ] 1

hybridation NC 0.001329 0.149 10 [ / 7523 / 67 ] 1

humanité NC 0.001196 0.225 9 [ / 7523 / 40 ] 1

immersion NC 0.001196 0.250 9 [ / 7523 / 36 ] 1

interopérabilité NC 0.001063 0.242 8 [ / 7523 / 33 ] 1

publicité NC 0.001063 0.190 8 [ / 7523 / 42 ] 1

promesse NC 0.001063 0.163 8 [ / 7523 / 49 ] 1

$URL$ NPP 0.001063 0.444 8 [ / 7523 / 18 ] 1

for domain 1.shs.droit we select 198 heads / 4166 ( 4.75 )

for domain 1.shs.droit we kept 112 heads / 4166 ( 2.69 )

lemma POS F / dom F / o occ [ / dom / occ

droit NC 0.041594 0.779 1098 [ / 26398 / 1410 ] 1 old

chronique NC 0.014509 0.696 383 [ / 26398 / 550 ] 1 old

responsabilité NC 0.010721 0.633 283 [ / 26398 / 447 ] 2 old

loi NC 0.008258 0.520 218 [ / 26398 / 419 ] 1 old

protection NC 0.008220 0.601 217 [ / 26398 / 361 ] 1 old

note NC 0.007917 0.239 209 [ / 26398 / 873 ] 1

réforme NC 0.006402 0.387 169 [ / 26398 / 437 ] 2

commentaire NC 0.006288 0.606 166 [ / 26398 / 274 ] 1 old

principe NC 0.005644 0.290 149 [ / 26398 / 514 ] 1

contrat NC 0.005341 0.613 141 [ / 26398 / 230 ] 1 old

obligation NC 0.005076 0.865 134 [ / 26398 / 155 ] 1 old

action NC 0.004962 0.240 131 [ / 26398 / 545 ] 2

juge NC 0.004887 0.822 129 [ / 26398 / 157 ] 1 old

contrôle NC 0.004584 0.182 121 [ / 26398 / 665 ] 3

notion NC 0.004546 0.270 120 [ / 26398 / 445 ] 1

régime NC 0.004053 0.423 107 [ / 26398 / 253 ] 1

liberté NC 0.004053 0.605 107 [ / 26398 / 177 ] 1 old

cour NC 0.004015 0.517 106 [ / 26398 / 205 ] 1 old

condition NC 0.003788 0.224 100 [ / 26398 / 447 ] 2

rapport NC 0.003750 0.196 99 [ / 26398 / 504 ] 2

procédure NC 0.003675 0.557 97 [ / 26398 / 174 ] 1 old

société NC 0.003561 0.279 94 [ / 26398 / 337 ] 2

assurance NC 0.003447 0.765 91 [ / 26398 / 119 ] 1 old

actualité NC 0.003371 0.405 89 [ / 26398 / 220 ] 1

union NC 0.003371 0.430 89 [ / 26398 / 207 ] 1

clause NC 0.003106 0.882 82 [ / 26398 / 93 ] 1 old

justice NC 0.003106 0.313 82 [ / 26398 / 262 ] 3

statut NC 0.003106 0.262 82 [ / 26398 / 313 ] 1

précision NC 0.002993 0.738 79 [ / 26398 / 107 ] 1 old

affaire NC 0.002841 0.328 75 [ / 26398 / 229 ] 2

accès NC 0.002765 0.285 73 [ / 26398 / 256 ] 2

article NC 0.002652 0.372 70 [ / 26398 / 188 ] 1

pouvoir NC 0.002576 0.210 68 [ / 26398 / 324 ] 2

accord NC 0.002538 0.549 67 [ / 26398 / 122 ] 1 old

compétence NC 0.002500 0.222 66 [ / 26398 / 297 ] 3

conseil NC 0.002462 0.489 65 [ / 26398 / 133 ] 1

règle NC 0.002387 0.391 63 [ / 26398 / 161 ] 1

contentieux NC 0.002273 0.909 60 [ / 26398 / 66 ] 1 old

reconnaissance NC 0.002273 0.186 60 [ / 26398 / 322 ] 2

conflit NC 0.002235 0.178 59 [ / 26398 / 332 ] 2

cadre NC 0.002235 0.159 59 [ / 26398 / 370 ] 2

propos NC 0.002197 0.580 58 [ / 26398 / 100 ] 1 old

preuve NC 0.002121 0.459 56 [ / 26398 / 122 ] 1

incidence NC 0.002083 0.279 55 [ / 26398 / 197 ] 2

arrêt NC 0.002083 0.764 55 [ / 26398 / 72 ] 1 old

sécurité NC 0.002008 0.301 53 [ / 26398 / 176 ] 2

convention NC 0.001970 0.505 52 [ / 26398 / 103 ] 1 old

fin NC 0.001970 0.193 52 [ / 26398 / 270 ] 2

règlement NC 0.001970 0.812 52 [ / 26398 / 64 ] 1 old

sanction NC 0.001894 0.746 50 [ / 26398 / 67 ] 1 old

recours NC 0.001894 0.410 50 [ / 26398 / 122 ] 1

jurisprudence NC 0.001856 0.860 49 [ / 26398 / 57 ] 1 old

limite NC 0.001856 0.167 49 [ / 26398 / 293 ] 1

participation NC 0.001856 0.205 49 [ / 26398 / 239 ] 2

ordre NC 0.001818 0.233 48 [ / 26398 / 206 ] 1

garantie NC 0.001780 0.979 47 [ / 26398 / 48 ] 1 old

prise NC 0.001780 0.228 47 [ / 26398 / 206 ] 2

autonomie NC 0.001705 0.312 45 [ / 26398 / 144 ] 1

entreprise NC 0.001667 0.171 44 [ / 26398 / 257 ] 3

égalité NC 0.001667 0.440 44 [ / 26398 / 100 ] 1

indemnisation NC 0.001667 0.786 44 [ / 26398 / 56 ] 1 old

cession NC 0.001667 0.936 44 [ / 26398 / 47 ] 1 old

constitution NC 0.001629 0.222 43 [ / 26398 / 194 ] 2

absence NC 0.001629 0.566 43 [ / 26398 / 76 ] 1 old

articulation NC 0.001629 0.269 43 [ / 26398 / 160 ] 1

panorama NC 0.001629 0.253 43 [ / 26398 / 170 ] 1

acte NC 0.001591 0.316 42 [ / 26398 / 133 ] 1

décision NC 0.001591 0.372 42 [ / 26398 / 113 ] 2

marché NC 0.001591 0.167 42 [ / 26398 / 251 ] 3

mariage NC 0.001591 0.353 42 [ / 26398 / 119 ] 1

avenir NC 0.001591 0.215 42 [ / 26398 / 195 ] 1

qualification NC 0.001591 0.525 42 [ / 26398 / 80 ] 1 old

arbitrage NC 0.001553 0.621 41 [ / 26398 / 66 ] 1 old

lutte NC 0.001553 0.205 41 [ / 26398 / 200 ] 2

code NC 0.001553 0.539 41 [ / 26398 / 76 ] 1 old

autorité NC 0.001515 0.357 40 [ / 26398 / 112 ] 1

nullité NC 0.001515 0.976 40 [ / 26398 / 41 ] 1 old

instrument NC 0.001440 0.230 38 [ / 26398 / 165 ] 1

fiscalité NC 0.001402 0.521 37 [ / 26398 / 71 ] 1 old

faute NC 0.001402 0.685 37 [ / 26398 / 54 ] 1 old

déclaration NC 0.001402 0.822 37 [ / 26398 / 45 ] 1 old

information NC 0.001402 0.185 37 [ / 26398 / 200 ] 2

veille NC 0.001402 0.446 37 [ / 26398 / 83 ] 1

France NPP 0.001364 0.169 36 [ / 26398 / 213 ] 3

financement NC 0.001364 0.259 36 [ / 26398 / 139 ] 2

force NC 0.001364 0.293 36 [ / 26398 / 123 ] 1

fondement NC 0.001288 0.195 34 [ / 26398 / 174 ] 2

refus NC 0.001288 0.531 34 [ / 26398 / 64 ] 1 old

encadrement NC 0.001288 0.576 34 [ / 26398 / 59 ] 1 old

spécificité NC 0.001288 0.174 34 [ / 26398 / 195 ] 1

rupture NC 0.001288 0.192 34 [ / 26398 / 177 ] 1

prescription NC 0.001288 0.500 34 [ / 26398 / 68 ] 1 old

démocratie NC 0.001288 0.179 34 [ / 26398 / 190 ] 2

sort NC 0.001288 0.618 34 [ / 26398 / 55 ] 1 old

doctrine NC 0.001250 0.508 33 [ / 26398 / 65 ] 1 old

situation NC 0.001250 0.158 33 [ / 26398 / 209 ] 2

disposition NC 0.001212 0.525 32 [ / 26398 / 61 ] 1 old

plan NC 0.001212 0.199 32 [ / 26398 / 161 ] 2

parlement NC 0.001174 0.492 31 [ / 26398 / 63 ] 2

réparation NC 0.001174 0.620 31 [ / 26398 / 50 ] 1 old

juridiction NC 0.001174 0.738 31 [ / 26398 / 42 ] 1 old

difficulté NC 0.001136 0.165 30 [ / 26398 / 182 ] 2

modification NC 0.001136 0.159 30 [ / 26398 / 189 ] 3

directive NC 0.001099 0.725 29 [ / 26398 / 40 ] 1 old

discrimination NC 0.001099 0.261 29 [ / 26398 / 111 ] 1

consécration NC 0.001061 0.651 28 [ / 26398 / 43 ] 1 old

collectivité NC 0.001061 0.549 28 [ / 26398 / 51 ] 1 old

santé NC 0.001061 0.174 28 [ / 26398 / 161 ] 1

réglementation NC 0.001061 0.549 28 [ / 26398 / 51 ] 1 old

exception NC 0.001023 0.482 27 [ / 26398 / 56 ] 1

plaidoyer NC 0.001023 0.290 27 [ / 26398 / 93 ] 1

décret NC 0.001023 0.794 27 [ / 26398 / 34 ] 1 old

for domain 1.shs.ling we select 185 heads / 3415 ( 5.42 )

for domain 1.shs.ling we kept 62 heads / 3415 ( 1.82 )

lemma POS F / dom F / o occ [ / dom / occ

langue NC 0.011153 0.474 173 [ / 15512 / 365 ] 1

discours NC 0.007220 0.243 112 [ / 15512 / 461 ] 1

expression NC 0.005802 0.330 90 [ / 15512 / 273 ] 1

mot NC 0.005222 0.312 81 [ / 15512 / 260 ] 1

nom NC 0.005157 0.465 80 [ / 15512 / 172 ] 1

traitement NC 0.004964 0.161 77 [ / 15512 / 478 ] 2

corpus NC 0.004577 0.493 71 [ / 15512 / 144 ] 2

traduction NC 0.004384 0.248 68 [ / 15512 / 274 ] 2

dictionnaire NC 0.003804 0.413 59 [ / 15512 / 143 ] 1

grammaire NC 0.003352 0.473 52 [ / 15512 / 110 ] 1

verbe NC 0.003352 0.722 52 [ / 15512 / 72 ] 1 old

français NC 0.003159 0.398 49 [ / 15512 / 123 ] 1

acquisition NC 0.003159 0.308 49 [ / 15512 / 159 ] 2

voix NC 0.002901 0.287 45 [ / 15512 / 157 ] 2

emploi NC 0.002837 0.195 44 [ / 15512 / 226 ] 3

didactique NC 0.002579 0.426 40 [ / 15512 / 94 ] 2

syntaxe NC 0.002514 0.661 39 [ / 15512 / 59 ] 1 old

langage NC 0.002514 0.207 39 [ / 15512 / 188 ] 2

terminologie NC 0.002321 0.720 36 [ / 15512 / 50 ] 1 old

marqueur NC 0.002256 0.333 35 [ / 15512 / 105 ] 1

linguistique NC 0.002256 0.686 35 [ / 15512 / 51 ] 1 old

phonologie NC 0.002192 0.829 34 [ / 15512 / 41 ] 1 old

base NC 0.002192 0.159 34 [ / 15512 / 214 ] 2

prosodie NC 0.002127 0.750 33 [ / 15512 / 44 ] 1 old

adjectif NC 0.002063 0.711 32 [ / 15512 / 45 ] 2 old

sémantique NC 0.001998 0.437 31 [ / 15512 / 71 ] 1

lexique NC 0.001934 0.435 30 [ / 15512 / 69 ] 1

grammaticalisation NC 0.001934 0.769 30 [ / 15512 / 39 ] 1 old

constitution NC 0.001870 0.149 29 [ / 15512 / 194 ] 2

contrainte NC 0.001741 0.155 27 [ / 15512 / 174 ] 1

préposition NC 0.001676 0.839 26 [ / 15512 / 31 ] 1 old

reformulation NC 0.001547 0.615 24 [ / 15512 / 39 ] 1 old

néologie NC 0.001483 1.000 23 [ / 15512 / 23 ] 1 old

figement NC 0.001418 0.917 22 [ / 15512 / 24 ] 1 old

référence NC 0.001354 0.231 21 [ / 15512 / 91 ] 1

phrase NC 0.001354 0.512 21 [ / 15512 / 41 ] 1 old

alternance NC 0.001354 0.553 21 [ / 15512 / 38 ] 1 old

polysémie NC 0.001354 0.636 21 [ / 15512 / 33 ] 1 old

type NC 0.001289 0.175 20 [ / 15512 / 114 ] 1

classe NC 0.001289 0.194 20 [ / 15512 / 103 ] 1

plurilinguisme NC 0.001289 0.667 20 [ / 15512 / 30 ] 1 old

emprunt NC 0.001289 0.370 20 [ / 15512 / 54 ] 1

négation NC 0.001289 0.588 20 [ / 15512 / 34 ] 1 old

contact NC 0.001225 0.268 19 [ / 15512 / 71 ] 1

énonciation NC 0.001225 0.543 19 [ / 15512 / 35 ] 1 old

désignation NC 0.001225 0.432 19 [ / 15512 / 44 ] 1

vocabulaire NC 0.001225 0.297 19 [ / 15512 / 64 ] 1

métaphore NC 0.001160 0.200 18 [ / 15512 / 90 ] 2

anaphore NC 0.001160 0.692 18 [ / 15512 / 26 ] 1 old

complément NC 0.001160 0.360 18 [ / 15512 / 50 ] 1

orthographe NC 0.001160 0.900 18 [ / 15512 / 20 ] 1 old

hétérogénéité NC 0.001160 0.250 18 [ / 15512 / 72 ] 1

particule NC 0.001096 0.515 17 [ / 15512 / 33 ] 1 old

dénomination NC 0.001096 0.586 17 [ / 15512 / 29 ] 1 old

marque NC 0.001096 0.172 17 [ / 15512 / 99 ] 1

repère NC 0.001096 0.185 17 [ / 15512 / 92 ] 1

pronom NC 0.001031 0.615 16 [ / 15512 / 26 ] 1 old

séquence NC 0.001031 0.242 16 [ / 15512 / 66 ] 1

liaison NC 0.001031 0.200 16 [ / 15512 / 80 ] 1

parler NC 0.001031 0.941 16 [ / 15512 / 17 ] 1 old

style NC 0.001031 0.188 16 [ / 15512 / 85 ] 1

locution NC 0.001031 0.842 16 [ / 15512 / 19 ] 1 old

for domain 1.shs.gestion we select 180 heads / 3529 ( 5.1 )

for domain 1.shs.gestion we kept 86 heads / 3529 ( 2.44 )

lemma POS F / dom F / o occ [ / dom / occ

cas NC 0.041495 0.238 1077 [ / 25955 / 4518 ] 1

analyse NC 0.024542 0.159 637 [ / 25955 / 4013 ] 2

rôle NC 0.012252 0.194 318 [ / 25955 / 1636 ] 1

impact NC 0.012098 0.243 314 [ / 25955 / 1290 ] 2

économie NC 0.008553 0.407 222 [ / 25955 / 546 ] 1

gestion NC 0.008438 0.226 219 [ / 25955 / 968 ] 1

stratégie NC 0.008207 0.227 213 [ / 25955 / 940 ] 1

proposition NC 0.007205 0.272 187 [ / 25955 / 688 ] 1

management NC 0.006935 0.669 180 [ / 25955 / 269 ] 1 old

déterminant NC 0.005856 0.551 152 [ / 25955 / 276 ] 1 old

dynamique NC 0.005163 0.156 134 [ / 25955 / 857 ] 2

gouvernance NC 0.004816 0.356 125 [ / 25955 / 351 ] 2

innovation NC 0.004816 0.321 125 [ / 25955 / 389 ] 1

perspective NC 0.004777 0.168 124 [ / 25955 / 737 ] 1

théorie NC 0.004700 0.182 122 [ / 25955 / 669 ] 2

contrôle NC 0.004392 0.171 114 [ / 25955 / 665 ] 3

crise NC 0.004277 0.276 111 [ / 25955 / 402 ] 1

entreprise NC 0.003930 0.397 102 [ / 25955 / 257 ] 3

intégration NC 0.003468 0.167 90 [ / 25955 / 538 ] 3

organisation NC 0.003468 0.197 90 [ / 25955 / 458 ] 1

responsabilité NC 0.003390 0.197 88 [ / 25955 / 447 ] 2

marché NC 0.003352 0.347 87 [ / 25955 / 251 ] 3

risque NC 0.003044 0.174 79 [ / 25955 / 455 ] 2

processus NC 0.003005 0.155 78 [ / 25955 / 502 ] 2

réforme NC 0.002928 0.174 76 [ / 25955 / 437 ] 2

émergence NC 0.002928 0.165 76 [ / 25955 / 460 ] 2

performance NC 0.002658 0.248 69 [ / 25955 / 278 ] 3

diversité NC 0.002620 0.189 68 [ / 25955 / 360 ] 2

revue NC 0.002620 0.221 68 [ / 25955 / 308 ] 2

concept NC 0.002581 0.165 67 [ / 25955 / 405 ] 1

régulation NC 0.002543 0.204 66 [ / 25955 / 324 ] 2

facteur NC 0.002504 0.188 65 [ / 25955 / 346 ] 2

comptabilité NC 0.002427 0.700 63 [ / 25955 / 90 ] 1 old

perception NC 0.002389 0.160 62 [ / 25955 / 387 ] 1

financement NC 0.002312 0.432 60 [ / 25955 / 139 ] 2

dimension NC 0.002312 0.183 60 [ / 25955 / 327 ] 2

défi NC 0.002312 0.151 60 [ / 25955 / 397 ] 2

leçon NC 0.002312 0.276 60 [ / 25955 / 217 ] 1

valeur NC 0.002273 0.168 59 [ / 25955 / 352 ] 2

croissance NC 0.002196 0.268 57 [ / 25955 / 213 ] 2

conséquence NC 0.002158 0.160 56 [ / 25955 / 350 ] 2

cadre NC 0.002158 0.151 56 [ / 25955 / 370 ] 2

communication NC 0.002158 0.179 56 [ / 25955 / 312 ] 2

création NC 0.002081 0.163 54 [ / 25955 / 332 ] 1

compétence NC 0.002003 0.175 52 [ / 25955 / 297 ] 3

investissement NC 0.002003 0.605 52 [ / 25955 / 86 ] 1 old

finance NC 0.001965 0.729 51 [ / 25955 / 70 ] 1 old

banque NC 0.001965 0.567 51 [ / 25955 / 90 ] 1 old

choix NC 0.001965 0.160 51 [ / 25955 / 319 ] 1

paradoxe NC 0.001888 0.186 49 [ / 25955 / 264 ] 2

logique NC 0.001811 0.151 47 [ / 25955 / 312 ] 2

qualité NC 0.001811 0.159 47 [ / 25955 / 296 ] 2

indicateur NC 0.001772 0.247 46 [ / 25955 / 186 ] 2

coopération NC 0.001734 0.215 45 [ / 25955 / 209 ] 1

coût NC 0.001695 0.419 44 [ / 25955 / 105 ] 1

concurrence NC 0.001657 0.371 43 [ / 25955 / 116 ] 1

information NC 0.001618 0.210 42 [ / 25955 / 200 ] 2

commerce NC 0.001618 0.253 42 [ / 25955 / 166 ] 2

marketing NC 0.001618 0.656 42 [ / 25955 / 64 ] 1 old

technologie NC 0.001618 0.165 42 [ / 25955 / 254 ] 1

fondement NC 0.001580 0.236 41 [ / 25955 / 174 ] 2

transport NC 0.001580 0.237 41 [ / 25955 / 173 ] 2

institution NC 0.001503 0.198 39 [ / 25955 / 197 ] 1

pilotage NC 0.001464 0.432 38 [ / 25955 / 88 ] 1

emploi NC 0.001464 0.168 38 [ / 25955 / 226 ] 3

éclairage NC 0.001464 0.264 38 [ / 25955 / 144 ] 1

illustration NC 0.001426 0.224 37 [ / 25955 / 165 ] 1

pôle NC 0.001387 0.507 36 [ / 25955 / 71 ] 1 old

efficacité NC 0.001348 0.183 35 [ / 25955 / 191 ] 2

capital NC 0.001348 0.417 35 [ / 25955 / 84 ] 1

industrie NC 0.001310 0.192 34 [ / 25955 / 177 ] 2

internationalisation NC 0.001310 0.540 34 [ / 25955 / 63 ] 1 old

mondialisation NC 0.001310 0.252 34 [ / 25955 / 135 ] 2

RSE NPP 0.001271 0.750 33 [ / 25955 / 44 ] 1 old

éthique NC 0.001233 0.225 32 [ / 25955 / 142 ] 2

décision NC 0.001194 0.274 31 [ / 25955 / 113 ] 2

pertinence NC 0.001156 0.259 30 [ / 25955 / 116 ] 1

difficulté NC 0.001156 0.165 30 [ / 25955 / 182 ] 2

monnaie NC 0.001117 0.227 29 [ / 25955 / 128 ] 3

inégalité NC 0.001117 0.180 29 [ / 25955 / 161 ] 2

importance NC 0.001040 0.161 27 [ / 25955 / 168 ] 2

coordination NC 0.001002 0.271 26 [ / 25955 / 96 ] 1

motivation NC 0.001002 0.295 26 [ / 25955 / 88 ] 1

opportunité NC 0.001002 0.286 26 [ / 25955 / 91 ] 1

normalisation NC 0.001002 0.260 26 [ / 25955 / 100 ] 1

confiance NC 0.001002 0.426 26 [ / 25955 / 61 ] 1

for domain 0.phys we select 140 heads / 3584 ( 3.91 )

for domain 0.phys we kept 100 heads / 3584 ( 2.79 )

lemma POS F / dom F / o occ [ / dom / occ

étude NC 0.080575 0.361 2471 [ / 30667 / 6842 ] 1

modélisation NC 0.035478 0.468 1088 [ / 30667 / 2324 ] 1

application NC 0.030293 0.368 929 [ / 30667 / 2525 ] 2

analyse NC 0.020185 0.154 619 [ / 30667 / 4013 ] 2

mesure NC 0.020022 0.513 614 [ / 30667 / 1198 ] 1 old

influence NC 0.019826 0.291 608 [ / 30667 / 2091 ] 2

méthode NC 0.018554 0.343 569 [ / 30667 / 1658 ] 2

caractérisation NC 0.017511 0.465 537 [ / 30667 / 1154 ] 1

effet NC 0.015195 0.210 466 [ / 30667 / 2221 ] 2

modèle NC 0.012000 0.193 368 [ / 30667 / 1910 ] 2

simulation NC 0.011804 0.537 362 [ / 30667 / 674 ] 1 old

détermination NC 0.009815 0.553 301 [ / 30667 / 544 ] 2 old

propriété NC 0.009554 0.539 293 [ / 30667 / 544 ] 1 old

conception NC 0.008152 0.299 250 [ / 30667 / 835 ] 2

évaluation NC 0.008087 0.154 248 [ / 30667 / 1608 ] 2

optimisation NC 0.007891 0.459 242 [ / 30667 / 527 ] 2

contribution NC 0.007793 0.197 239 [ / 30667 / 1214 ] 1

recherche NC 0.007728 0.187 237 [ / 30667 / 1268 ] 2

utilisation NC 0.007109 0.185 218 [ / 30667 / 1181 ] 2

comparaison NC 0.006717 0.199 206 [ / 30667 / 1036 ] 2

comportement NC 0.006293 0.398 193 [ / 30667 / 485 ] 2

calcul NC 0.006261 0.535 192 [ / 30667 / 359 ] 2 old

identification NC 0.005935 0.367 182 [ / 30667 / 496 ] 1

développement NC 0.005902 0.170 181 [ / 30667 / 1065 ] 1

diffusion NC 0.005609 0.444 172 [ / 30667 / 387 ] 1

spectre NC 0.005380 0.846 165 [ / 30667 / 195 ] 1 old

estimation NC 0.005348 0.283 164 [ / 30667 / 579 ] 4

structure NC 0.005315 0.265 163 [ / 30667 / 614 ] 1

théorie NC 0.005120 0.235 157 [ / 30667 / 669 ] 2

détection NC 0.004598 0.294 141 [ / 30667 / 479 ] 2

réalisation NC 0.004435 0.567 136 [ / 30667 / 240 ] 2 old

contrôle NC 0.004337 0.200 133 [ / 30667 / 665 ] 3

interaction NC 0.004076 0.244 125 [ / 30667 / 513 ] 1

remarque NC 0.003978 0.200 122 [ / 30667 / 611 ] 1

synthèse NC 0.003489 0.189 107 [ / 30667 / 565 ] 2

antenne NC 0.003391 0.680 104 [ / 30667 / 153 ] 2 old

dispositif NC 0.003359 0.186 103 [ / 30667 / 553 ] 1

observation NC 0.003293 0.213 101 [ / 30667 / 475 ] 2

commande NC 0.003261 0.602 100 [ / 30667 / 166 ] 1 old

couplage NC 0.003163 0.595 97 [ / 30667 / 163 ] 1 old

propagation NC 0.002870 0.698 88 [ / 30667 / 126 ] 1 old

amélioration NC 0.002870 0.342 88 [ / 30667 / 257 ] 3

mécanisme NC 0.002837 0.290 87 [ / 30667 / 300 ] 2

technique NC 0.002804 0.232 86 [ / 30667 / 370 ] 1

génération NC 0.002674 0.279 82 [ / 30667 / 294 ] 2

capteur NC 0.002641 0.675 81 [ / 30667 / 120 ] 1 old

intégration NC 0.002609 0.149 80 [ / 30667 / 538 ] 3

méthodologie NC 0.002511 0.208 77 [ / 30667 / 370 ] 1

appareil NC 0.002380 0.716 73 [ / 30667 / 102 ] 1 old

transition NC 0.002315 0.270 71 [ / 30667 / 263 ] 2

traitement NC 0.002315 0.149 71 [ / 30667 / 478 ] 2

émission NC 0.002315 0.526 71 [ / 30667 / 135 ] 2 old

spectroscopie NC 0.002250 0.663 69 [ / 30667 / 104 ] 2 old

résonance NC 0.002217 0.756 68 [ / 30667 / 90 ] 1 old

absorption NC 0.002152 0.673 66 [ / 30667 / 98 ] 2 old

matériau NC 0.002087 0.344 64 [ / 30667 / 186 ] 3

phénomène NC 0.001989 0.305 61 [ / 30667 / 200 ] 1

diagnostic NC 0.001989 0.271 61 [ / 30667 / 225 ] 2

description NC 0.001924 0.230 59 [ / 30667 / 256 ] 1

élaboration NC 0.001891 0.197 58 [ / 30667 / 294 ] 1

interprétation NC 0.001891 0.188 58 [ / 30667 / 308 ] 1

réduction NC 0.001826 0.303 56 [ / 30667 / 185 ] 1

laser NC 0.001793 0.821 55 [ / 30667 / 67 ] 1 old

instabilité NC 0.001761 0.643 54 [ / 30667 / 84 ] 1 old

formulation NC 0.001728 0.541 53 [ / 30667 / 98 ] 1 old

source NC 0.001728 0.149 53 [ / 30667 / 355 ] 2

transfert NC 0.001696 0.191 52 [ / 30667 / 272 ] 1

performance NC 0.001696 0.187 52 [ / 30667 / 278 ] 3

imagerie NC 0.001663 0.354 51 [ / 30667 / 144 ] 2

validation NC 0.001663 0.245 51 [ / 30667 / 208 ] 3

onde NC 0.001630 0.794 50 [ / 30667 / 63 ] 1 old

prédiction NC 0.001565 0.356 48 [ / 30667 / 135 ] 3

corrélation NC 0.001565 0.545 48 [ / 30667 / 88 ] 1 old

écoulement NC 0.001533 0.825 47 [ / 30667 / 57 ] 1 old

niveau NC 0.001533 0.392 47 [ / 30667 / 120 ] 1

énergie NC 0.001500 0.451 46 [ / 30667 / 102 ] 1

modification NC 0.001500 0.243 46 [ / 30667 / 189 ] 3

solution NC 0.001435 0.265 44 [ / 30667 / 166 ] 2

diffraction NC 0.001435 0.880 44 [ / 30667 / 50 ] 1 old

réaction NC 0.001370 0.356 42 [ / 30667 / 118 ] 1

excitation NC 0.001370 0.955 42 [ / 30667 / 44 ] 1 old

filtre NC 0.001370 0.618 42 [ / 30667 / 68 ] 1 old

couche NC 0.001337 0.651 41 [ / 30667 / 63 ] 1 old

stabilité NC 0.001337 0.410 41 [ / 30667 / 100 ] 1

distribution NC 0.001304 0.252 40 [ / 30667 / 159 ] 1

extension NC 0.001272 0.179 39 [ / 30667 / 218 ] 2

micro NC 0.001207 0.339 37 [ / 30667 / 109 ] 1

caractéristique NC 0.001207 0.175 37 [ / 30667 / 212 ] 2

dimensionnement NC 0.001207 0.673 37 [ / 30667 / 55 ] 1 old

localisation NC 0.001207 0.268 37 [ / 30667 / 138 ] 2

mouvement NC 0.001174 0.150 36 [ / 30667 / 240 ] 1

champ NC 0.001174 0.222 36 [ / 30667 / 162 ] 1

résolution NC 0.001174 0.229 36 [ / 30667 / 157 ] 2

polarisation NC 0.001174 0.706 36 [ / 30667 / 51 ] 1 old

équation NC 0.001141 0.538 35 [ / 30667 / 65 ] 1 old

expérimentation NC 0.001141 0.168 35 [ / 30667 / 208 ] 1

rayonnement NC 0.001076 0.647 33 [ / 30667 / 51 ] 1 old

préparation NC 0.001076 0.311 33 [ / 30667 / 106 ] 2

relaxation NC 0.001043 0.762 32 [ / 30667 / 42 ] 1 old

prévision NC 0.001043 0.311 32 [ / 30667 / 103 ] 1

for domain 1.shs.anthro we select 197 heads / 2573 ( 7.66 )

for domain 1.shs.anthro we kept 20 heads / 2573 ( 0.78 )

lemma POS F / dom F / o occ [ / dom / occ

anthropologie NC 0.009795 0.420 68 [ / 6942 / 162 ] 1

ethnographie NC 0.004177 0.305 29 [ / 6942 / 95 ] 1

ethnologie NC 0.002593 0.750 18 [ / 6942 / 24 ] 1 old

rituel NC 0.002593 0.300 18 [ / 6942 / 60 ] 1

rite NC 0.002449 0.283 17 [ / 6942 / 60 ] 1

population NC 0.002449 0.179 17 [ / 6942 / 95 ] 2

médecine NC 0.002161 0.192 15 [ / 6942 / 78 ] 1

dieu NC 0.002017 0.259 14 [ / 6942 / 54 ] 1

ethnologue NC 0.001873 0.765 13 [ / 6942 / 17 ] 1 old

terrain NC 0.001729 0.152 12 [ / 6942 / 79 ] 1

parenté NC 0.001585 0.333 11 [ / 6942 / 33 ] 1

culte NC 0.001441 0.152 10 [ / 6942 / 66 ] 2

Népal NPP 0.001152 0.889 8 [ / 6942 / 9 ] 1 old

cimetière NC 0.001152 0.258 8 [ / 6942 / 31 ] 1

pentecôtisme NC 0.001152 0.500 8 [ / 6942 / 16 ] 1 old

collecte NC 0.001008 0.175 7 [ / 6942 / 40 ] 1

ancêtre NC 0.001008 0.500 7 [ / 6942 / 14 ] 1 old

sacrifice NC 0.001008 0.259 7 [ / 6942 / 27 ] 1

anthropologue NC 0.001008 0.500 7 [ / 6942 / 14 ] 1 old

émeute NC 0.001008 0.219 7 [ / 6942 / 32 ] 1

for domain 1.shs.hist we select 163 heads / 6984 ( 2.33 )

for domain 1.shs.hist we kept 70 heads / 6984 ( 1.0 )

lemma POS F / dom F / o occ [ / dom / occ

histoire NC 0.012699 0.236 326 [ / 25671 / 1383 ] 1

image NC 0.004441 0.200 114 [ / 25671 / 570 ] 1

femme NC 0.003934 0.263 101 [ / 25671 / 384 ] 2

mémoire NC 0.003779 0.246 97 [ / 25671 / 394 ] 1

compte NC 0.003467 0.186 89 [ / 25671 / 478 ] 1

source NC 0.003116 0.225 80 [ / 25671 / 355 ] 2

société NC 0.003077 0.234 79 [ / 25671 / 337 ] 2

guerre NC 0.002961 0.317 76 [ / 25671 / 240 ] 1

livre NC 0.002922 0.330 75 [ / 25671 / 227 ] 2

autour NC 0.002844 0.232 73 [ / 25671 / 315 ] 1

inscription NC 0.002805 0.343 72 [ / 25671 / 210 ] 2

notice NC 0.002571 0.219 66 [ / 25671 / 302 ] 3

identité NC 0.002571 0.152 66 [ / 25671 / 435 ] 2

archive NC 0.002493 0.325 64 [ / 25671 / 197 ] 1

enquête NC 0.002493 0.148 64 [ / 25671 / 431 ] 2

voyage NC 0.002376 0.250 61 [ / 25671 / 244 ] 2

vie NC 0.002337 0.200 60 [ / 25671 / 300 ] 1

origine NC 0.002298 0.166 59 [ / 25671 / 355 ] 1

lieu NC 0.002220 0.200 57 [ / 25671 / 285 ] 2

naissance NC 0.002220 0.226 57 [ / 25671 / 252 ] 1

église NC 0.002104 0.310 54 [ / 25671 / 174 ] 3

musique NC 0.002026 0.198 52 [ / 25671 / 262 ] 2

récit NC 0.001987 0.173 51 [ / 25671 / 295 ] 2

manuscrit NC 0.001987 0.455 51 [ / 25671 / 112 ] 3

témoignage NC 0.001987 0.315 51 [ / 25671 / 162 ] 2

révolution NC 0.001948 0.230 50 [ / 25671 / 217 ] 1

pouvoir NC 0.001948 0.154 50 [ / 25671 / 324 ] 2

historiographie NC 0.001792 0.719 46 [ / 25671 / 64 ] 1 old

affaire NC 0.001792 0.201 46 [ / 25671 / 229 ] 2

mythe NC 0.001753 0.180 45 [ / 25671 / 250 ] 1

religion NC 0.001753 0.290 45 [ / 25671 / 155 ] 1

France NPP 0.001714 0.207 44 [ / 25671 / 213 ] 3

homme NC 0.001714 0.169 44 [ / 25671 / 260 ] 2

justice NC 0.001714 0.168 44 [ / 25671 / 262 ] 3

débat NC 0.001714 0.176 44 [ / 25671 / 250 ] 1

circulation NC 0.001675 0.184 43 [ / 25671 / 234 ] 1

archéologie NC 0.001636 0.179 42 [ / 25671 / 234 ] 2

famille NC 0.001636 0.194 42 [ / 25671 / 216 ] 2

monde NC 0.001636 0.179 42 [ / 25671 / 234 ] 2

lettre NC 0.001636 0.266 42 [ / 25671 / 158 ] 2

siècle NC 0.001558 0.244 40 [ / 25671 / 164 ] 1

bibliothèque NC 0.001558 0.247 40 [ / 25671 / 162 ] 2

évêque NC 0.001519 0.867 39 [ / 25671 / 45 ] 1 old

présence NC 0.001519 0.214 39 [ / 25671 / 182 ] 2

noblesse NC 0.001519 0.765 39 [ / 25671 / 51 ] 1 old

élite NC 0.001519 0.411 39 [ / 25671 / 95 ] 1

communauté NC 0.001519 0.181 39 [ / 25671 / 216 ] 1

mort NC 0.001480 0.266 38 [ / 25671 / 143 ] 2

portrait NC 0.001441 0.178 37 [ / 25671 / 208 ] 3

face NC 0.001441 0.224 37 [ / 25671 / 165 ] 1

invention NC 0.001441 0.167 37 [ / 25671 / 221 ] 2

château NC 0.001441 0.301 37 [ / 25671 / 123 ] 2

fragment NC 0.001363 0.285 35 [ / 25671 / 123 ] 2

texte NC 0.001324 0.211 34 [ / 25671 / 161 ] 2

tradition NC 0.001324 0.238 34 [ / 25671 / 143 ] 1

école NC 0.001285 0.156 33 [ / 25671 / 212 ] 3

itinéraire NC 0.001285 0.223 33 [ / 25671 / 148 ] 1

héritage NC 0.001247 0.167 32 [ / 25671 / 192 ] 1

maison NC 0.001208 0.196 31 [ / 25671 / 158 ] 2

édition NC 0.001208 0.226 31 [ / 25671 / 137 ] 2

État NC 0.001208 0.183 31 [ / 25671 / 169 ] 2

monnaie NC 0.001130 0.227 29 [ / 25671 / 128 ] 3

culte NC 0.001130 0.439 29 [ / 25671 / 66 ] 2

procès NC 0.001130 0.341 29 [ / 25671 / 85 ] 1

début NC 0.001091 0.269 28 [ / 25671 / 104 ] 1

Paris NPP 0.001052 0.235 27 [ / 25671 / 115 ] 2

Rome NPP 0.001052 0.540 27 [ / 25671 / 50 ] 1 old

Jean NPP 0.001013 0.211 26 [ / 25671 / 123 ] 2

traité NC 0.001013 0.325 26 [ / 25671 / 80 ] 1

historien NC 0.001013 0.500 26 [ / 25671 / 52 ] 1 old

for domain 0.sde we select 183 heads / 1971 ( 9.28 )

for domain 0.sde we kept 22 heads / 1971 ( 1.12 )

lemma POS F / dom F / o occ [ / dom / occ

bibliographie NC 0.008418 0.453 63 [ / 7484 / 139 ] 1

brève NC 0.008418 0.984 63 [ / 7484 / 64 ] 1 old

agriculture NC 0.005612 0.194 42 [ / 7484 / 216 ] 3

valorisation NC 0.003741 0.156 28 [ / 7484 / 179 ] 2

émission NC 0.002940 0.163 22 [ / 7484 / 135 ] 2

écologie NC 0.002539 0.161 19 [ / 7484 / 118 ] 2

observatoire NC 0.002539 0.190 19 [ / 7484 / 100 ] 1

forêt NC 0.002405 0.207 18 [ / 7484 / 87 ] 1

restauration NC 0.002272 0.202 17 [ / 7484 / 84 ] 2

biodiversité NC 0.002004 0.163 15 [ / 7484 / 92 ] 2

karst NC 0.001871 0.560 14 [ / 7484 / 25 ] 2 old

pollution NC 0.001737 0.191 13 [ / 7484 / 68 ] 1

recension NC 0.001737 0.232 13 [ / 7484 / 56 ] 1

muraille NC 0.001603 0.600 12 [ / 7484 / 20 ] 1 old

clé NC 0.001603 0.150 12 [ / 7484 / 80 ] 1

sol NC 0.001470 0.208 11 [ / 7484 / 53 ] 1

espèce NC 0.001470 0.262 11 [ / 7484 / 42 ] 1

contamination NC 0.001470 0.282 11 [ / 7484 / 39 ] 1

parc NC 0.001203 0.153 9 [ / 7484 / 59 ] 1

cadastre NC 0.001203 0.562 9 [ / 7484 / 16 ] 1 old

invasion NC 0.001069 0.400 8 [ / 7484 / 20 ] 1

focus NC 0.001069 0.229 8 [ / 7484 / 35 ] 1

for domain 1.shs.phil we select 199 heads / 2797 ( 7.11 )

for domain 1.shs.phil we kept 31 heads / 2797 ( 1.11 )

lemma POS F / dom F / o occ [ / dom / occ

philosophie NC 0.010438 0.482 82 [ / 7856 / 170 ] 1

science NC 0.008656 0.159 68 [ / 7856 / 428 ] 2

critique NC 0.006746 0.209 53 [ / 7856 / 253 ] 1

idée NC 0.003946 0.205 31 [ / 7856 / 151 ] 1

éthique NC 0.003310 0.183 26 [ / 7856 / 142 ] 2

épistémologie NC 0.003182 0.321 25 [ / 7856 / 78 ] 1

journée NC 0.002928 0.205 23 [ / 7856 / 112 ] 1

colloque NC 0.002546 0.238 20 [ / 7856 / 84 ] 1

raison NC 0.002419 0.165 19 [ / 7856 / 115 ] 1

mathématique NC 0.002419 0.253 19 [ / 7856 / 75 ] 3

Leibniz NPP 0.002037 0.941 16 [ / 7856 / 17 ] 1 old

philosophe NC 0.001909 0.484 15 [ / 7856 / 31 ] 1

Spinoza NPP 0.001909 0.714 15 [ / 7856 / 21 ] 1 old

Descartes NPP 0.001782 0.778 14 [ / 7856 / 18 ] 1 old

Kant NPP 0.001782 0.778 14 [ / 7856 / 18 ] 1 old

Nietzsche NPP 0.001782 0.560 14 [ / 7856 / 25 ] 1 old

Foucault NPP 0.001655 0.500 13 [ / 7856 / 26 ] 1 old

Bergson NPP 0.001527 0.857 12 [ / 7856 / 14 ] 1 old

rationalité NC 0.001527 0.226 12 [ / 7856 / 53 ] 1

Diderot NPP 0.001527 0.414 12 [ / 7856 / 29 ] 1

réalisme NC 0.001400 0.250 11 [ / 7856 / 44 ] 2

explication NC 0.001400 0.196 11 [ / 7856 / 56 ] 1

vérité NC 0.001400 0.193 11 [ / 7856 / 57 ] 1

Henri Poincaré NPP 0.001400 0.550 11 [ / 7856 / 20 ] 2 old

scepticisme NC 0.001400 0.550 11 [ / 7856 / 20 ] 1 old

Poincaré NPP 0.001273 0.769 10 [ / 7856 / 13 ] 1 old

imagination NC 0.001146 0.310 9 [ / 7856 / 29 ] 1

Aristote NPP 0.001146 0.391 9 [ / 7856 / 23 ] 1

intuition NC 0.001146 0.600 9 [ / 7856 / 15 ] 1 old

phénoménologie NC 0.001018 0.364 8 [ / 7856 / 22 ] 1

Harbermas NPP 0.001018 1.000 8 [ / 7856 / 8 ] 1 old

for domain NONE we select 167 heads / 5896 ( 2.83 )

for domain NONE we kept 6 heads / 5896 ( 0.1 )

lemma POS F / dom F / o occ [ / dom / occ

art NC 0.005140 0.226 127 [ / 24708 / 563 ] 2

réalisation NC 0.001619 0.167 40 [ / 24708 / 240 ] 2

antenne NC 0.001578 0.255 39 [ / 24708 / 153 ] 2

imagerie NC 0.001417 0.243 35 [ / 24708 / 144 ] 2

matériau NC 0.001295 0.172 32 [ / 24708 / 186 ] 3

plan NC 0.001012 0.155 25 [ / 24708 / 161 ] 2

for domain 0.sdv we select 157 heads / 3792 ( 4.14 )

for domain 0.sdv we kept 71 heads / 3792 ( 1.87 )

lemma POS F / dom F / o occ [ / dom / occ

influence NC 0.031830 0.337 705 [ / 22149 / 2091 ] 2

effet NC 0.031830 0.317 705 [ / 22149 / 2221 ] 2

utilisation NC 0.012687 0.238 281 [ / 22149 / 1181 ] 2

évolution NC 0.012506 0.160 277 [ / 22149 / 1730 ] 1

évaluation NC 0.011784 0.162 261 [ / 22149 / 1608 ] 2

impact NC 0.010294 0.177 228 [ / 22149 / 1290 ] 2

comparaison NC 0.010113 0.216 224 [ / 22149 / 1036 ] 2

recherche NC 0.009391 0.164 208 [ / 22149 / 1268 ] 2

relation NC 0.007495 0.158 166 [ / 22149 / 1051 ] 1

intérêt NC 0.006682 0.299 148 [ / 22149 / 495 ] 1

résultat NC 0.005869 0.212 130 [ / 22149 / 613 ] 2

variation NC 0.005689 0.249 126 [ / 22149 / 506 ] 1

production NC 0.004695 0.192 104 [ / 22149 / 543 ] 2

facteur NC 0.004425 0.283 98 [ / 22149 / 346 ] 2

conséquence NC 0.004289 0.271 95 [ / 22149 / 350 ] 2

estimation NC 0.004018 0.154 89 [ / 22149 / 579 ] 4

action NC 0.003747 0.152 83 [ / 22149 / 545 ] 2

détermination NC 0.003657 0.149 81 [ / 22149 / 544 ] 2

valeur NC 0.003657 0.230 81 [ / 22149 / 352 ] 2

dosage NC 0.003612 0.800 80 [ / 22149 / 100 ] 1 old

qualité NC 0.003431 0.257 76 [ / 22149 / 296 ] 2

comportement NC 0.003386 0.155 75 [ / 22149 / 485 ] 2

activité NC 0.003296 0.195 73 [ / 22149 / 375 ] 1

diversité NC 0.003296 0.203 73 [ / 22149 / 360 ] 2

observation NC 0.003251 0.152 72 [ / 22149 / 475 ] 2

composition NC 0.003160 0.424 70 [ / 22149 / 165 ] 1

modification NC 0.002935 0.344 65 [ / 22149 / 189 ] 3

alimentation NC 0.002799 0.446 62 [ / 22149 / 139 ] 1

variabilité NC 0.002754 0.324 61 [ / 22149 / 188 ] 3

adaptation NC 0.002754 0.181 61 [ / 22149 / 337 ] 2

régulation NC 0.002709 0.185 60 [ / 22149 / 324 ] 2

agriculture NC 0.002528 0.259 56 [ / 22149 / 216 ] 3

croissance NC 0.002483 0.258 55 [ / 22149 / 213 ] 2

importance NC 0.002438 0.321 54 [ / 22149 / 168 ] 2

mécanisme NC 0.002438 0.180 54 [ / 22149 / 300 ] 2

sélection NC 0.002348 0.306 52 [ / 22149 / 170 ] 2

caractéristique NC 0.002348 0.245 52 [ / 22149 / 212 ] 2

revue NC 0.002348 0.169 52 [ / 22149 / 308 ] 2

implication NC 0.002257 0.196 50 [ / 22149 / 255 ] 1

diagnostic NC 0.002167 0.213 48 [ / 22149 / 225 ] 2

emploi NC 0.002167 0.212 48 [ / 22149 / 226 ] 3

efficacité NC 0.002077 0.241 46 [ / 22149 / 191 ] 2

lait NC 0.002032 0.865 45 [ / 22149 / 52 ] 1 old

réponse NC 0.002032 0.163 45 [ / 22149 / 276 ] 1

résistance NC 0.002032 0.209 45 [ / 22149 / 215 ] 1

performance NC 0.001987 0.158 44 [ / 22149 / 278 ] 3

incidence NC 0.001987 0.223 44 [ / 22149 / 197 ] 2

amélioration NC 0.001941 0.167 43 [ / 22149 / 257 ] 3

maladie NC 0.001851 0.436 41 [ / 22149 / 94 ] 1

base NC 0.001806 0.187 40 [ / 22149 / 214 ] 2

validation NC 0.001671 0.178 37 [ / 22149 / 208 ] 3

élevage NC 0.001625 0.581 36 [ / 22149 / 62 ] 1 old

possibilité NC 0.001625 0.336 36 [ / 22149 / 107 ] 1

exposition NC 0.001535 0.260 34 [ / 22149 / 131 ] 1

conservation NC 0.001445 0.390 32 [ / 22149 / 82 ] 1

association NC 0.001400 0.162 31 [ / 22149 / 191 ] 2

protéine NC 0.001354 0.909 30 [ / 22149 / 33 ] 1 old

aptitude NC 0.001354 0.811 30 [ / 22149 / 37 ] 1 old

paramètre NC 0.001309 0.382 29 [ / 22149 / 76 ] 1

prédiction NC 0.001309 0.215 29 [ / 22149 / 135 ] 3

indicateur NC 0.001309 0.156 29 [ / 22149 / 186 ] 2

valorisation NC 0.001264 0.156 28 [ / 22149 / 179 ] 2

absorption NC 0.001219 0.276 27 [ / 22149 / 98 ] 2

épidémiologie NC 0.001219 0.587 27 [ / 22149 / 46 ] 1 old

prévalence NC 0.001219 0.750 27 [ / 22149 / 36 ] 1 old

déterminisme NC 0.001219 0.659 27 [ / 22149 / 41 ] 1 old

besoin NC 0.001129 0.362 25 [ / 22149 / 69 ] 1

écologie NC 0.001129 0.212 25 [ / 22149 / 118 ] 2

cinétique NC 0.001129 0.417 25 [ / 22149 / 60 ] 1

isolement NC 0.001084 0.686 24 [ / 22149 / 35 ] 1 old

biodiversité NC 0.001084 0.261 24 [ / 22149 / 92 ] 2

for domain 1.shs.archi we select 194 heads / 1614 ( 12.02 )

for domain 1.shs.archi we kept 16 heads / 1614 ( 0.99 )

lemma POS F / dom F / o occ [ / dom / occ

architecture NC 0.014690 0.155 68 [ / 4629 / 439 ] 2

ambiance NC 0.010801 0.794 50 [ / 4629 / 63 ] 1 old

urbanisme NC 0.007777 0.500 36 [ / 4629 / 72 ] 1 old

Paris NPP 0.003889 0.157 18 [ / 4629 / 115 ] 2

quartier NC 0.003240 0.149 15 [ / 4629 / 101 ] 3

fortification NC 0.003024 0.246 14 [ / 4629 / 57 ] 2

architecte NC 0.002376 0.324 11 [ / 4629 / 34 ] 2

logement NC 0.002376 0.244 11 [ / 4629 / 45 ] 1

patrimonialisation NC 0.002160 0.175 10 [ / 4629 / 57 ] 1

photogrammétrie NC 0.001944 0.643 9 [ / 4629 / 14 ] 1 old

métropole NC 0.001944 0.164 9 [ / 4629 / 55 ] 2

marche NC 0.001728 0.148 8 [ / 4629 / 54 ] 1

rue NC 0.001512 0.159 7 [ / 4629 / 44 ] 1

banlieue NC 0.001296 0.250 6 [ / 4629 / 24 ] 1

urbanité NC 0.001080 0.227 5 [ / 4629 / 22 ] 1

habitant NC 0.001080 0.161 5 [ / 4629 / 31 ] 1

for domain 0.info we select 157 heads / 3273 ( 4.8 )

for domain 0.info we kept 74 heads / 3273 ( 2.26 )

lemma POS F / dom F / o occ [ / dom / occ

application NC 0.024937 0.160 405 [ / 16241 / 2525 ] 2

modèle NC 0.022905 0.195 372 [ / 16241 / 1910 ] 2

méthode NC 0.015639 0.153 254 [ / 16241 / 1658 ] 2

algorithme NC 0.014039 0.683 228 [ / 16241 / 334 ] 1 old

système NC 0.013915 0.211 226 [ / 16241 / 1072 ] 1

extraction NC 0.010344 0.598 168 [ / 16241 / 281 ] 1 old

détection NC 0.010283 0.349 167 [ / 16241 / 479 ] 2

outil NC 0.009852 0.153 160 [ / 16241 / 1045 ] 1

apprentissage NC 0.009051 0.340 147 [ / 16241 / 432 ] 1

conception NC 0.008312 0.162 135 [ / 16241 / 835 ] 2

optimisation NC 0.007697 0.237 125 [ / 16241 / 527 ] 2

architecture NC 0.007081 0.262 115 [ / 16241 / 439 ] 2

segmentation NC 0.006958 0.614 113 [ / 16241 / 184 ] 1 old

classification NC 0.006404 0.390 104 [ / 16241 / 267 ] 1

intégration NC 0.005726 0.173 93 [ / 16241 / 538 ] 3

génération NC 0.005726 0.316 93 [ / 16241 / 294 ] 2

estimation NC 0.005665 0.159 92 [ / 16241 / 579 ] 4

reconnaissance NC 0.004926 0.248 80 [ / 16241 / 322 ] 2

ordonnancement NC 0.004926 0.879 80 [ / 16241 / 91 ] 1 old

visualisation NC 0.003571 0.569 58 [ / 16241 / 102 ] 1 old

calcul NC 0.003571 0.162 58 [ / 16241 / 359 ] 2

reconstruction NC 0.003510 0.358 57 [ / 16241 / 159 ] 1

planification NC 0.003448 0.431 56 [ / 16241 / 130 ] 1

adaptation NC 0.003386 0.163 55 [ / 16241 / 337 ] 2

extension NC 0.003325 0.248 54 [ / 16241 / 218 ] 2

résolution NC 0.003263 0.338 53 [ / 16241 / 157 ] 2

amélioration NC 0.003202 0.202 52 [ / 16241 / 257 ] 3

plateforme NC 0.003140 0.364 51 [ / 16241 / 140 ] 1

langage NC 0.002955 0.255 48 [ / 16241 / 188 ] 2

protocole NC 0.002955 0.432 48 [ / 16241 / 111 ] 1

programmation NC 0.002832 0.605 46 [ / 16241 / 76 ] 1 old

ontologie NC 0.002832 0.561 46 [ / 16241 / 82 ] 1 old

environnement NC 0.002771 0.158 45 [ / 16241 / 284 ] 1

interface NC 0.002709 0.400 44 [ / 16241 / 110 ] 1

fouille NC 0.002648 0.247 43 [ / 16241 / 174 ] 2

annotation NC 0.002648 0.551 43 [ / 16241 / 78 ] 1 old

fusion NC 0.002586 0.389 42 [ / 16241 / 108 ] 1

plate NC 0.002524 0.441 41 [ / 16241 / 93 ] 1

exploration NC 0.002463 0.196 40 [ / 16241 / 204 ] 1

aide NC 0.002463 0.172 40 [ / 16241 / 232 ] 1

indexation NC 0.002401 0.736 39 [ / 16241 / 53 ] 1 old

vérification NC 0.002278 0.500 37 [ / 16241 / 74 ] 1 old

validation NC 0.002278 0.178 37 [ / 16241 / 208 ] 3

combinaison NC 0.002217 0.404 36 [ / 16241 / 89 ] 1

heuristique NC 0.002155 0.714 35 [ / 16241 / 49 ] 1 old

sélection NC 0.002032 0.194 33 [ / 16241 / 170 ] 2

formalisation NC 0.001970 0.416 32 [ / 16241 / 77 ] 1

routage NC 0.001847 0.938 30 [ / 16241 / 32 ] 1 old

logiciel NC 0.001786 0.248 29 [ / 16241 / 117 ] 1

spécification NC 0.001786 0.690 29 [ / 16241 / 42 ] 1 old

sécurité NC 0.001724 0.159 28 [ / 16241 / 176 ] 2

prédiction NC 0.001662 0.200 27 [ / 16241 / 135 ] 3

solution NC 0.001539 0.151 25 [ / 16241 / 166 ] 2

intelligence NC 0.001539 0.194 25 [ / 16241 / 129 ] 2

minimisation NC 0.001539 0.758 25 [ / 16241 / 33 ] 1 old

acquisition NC 0.001478 0.151 24 [ / 16241 / 159 ] 2

raisonnement NC 0.001478 0.500 24 [ / 16241 / 48 ] 1 old

décomposition NC 0.001416 0.324 23 [ / 16241 / 71 ] 1

graphe NC 0.001416 0.535 23 [ / 16241 / 43 ] 1 old

corpus NC 0.001416 0.160 23 [ / 16241 / 144 ] 2

auto NC 0.001416 0.202 23 [ / 16241 / 114 ] 1

approximation NC 0.001416 0.348 23 [ / 16241 / 66 ] 2

ingénierie NC 0.001355 0.171 22 [ / 16241 / 129 ] 1

filtrage NC 0.001355 0.355 22 [ / 16241 / 62 ] 1

schéma NC 0.001293 0.200 21 [ / 16241 / 105 ] 1

localisation NC 0.001293 0.152 21 [ / 16241 / 138 ] 2

implémentation NC 0.001293 0.362 21 [ / 16241 / 58 ] 1

compression NC 0.001231 0.435 20 [ / 16241 / 46 ] 1

enrichissement NC 0.001231 0.417 20 [ / 16241 / 48 ] 1

alignement NC 0.001108 0.409 18 [ / 16241 / 44 ] 1

complexité NC 0.001047 0.173 17 [ / 16241 / 98 ] 1

recalage NC 0.001047 0.472 17 [ / 16241 / 36 ] 1

séparation NC 0.001047 0.189 17 [ / 16241 / 90 ] 1

borne NC 0.001047 0.425 17 [ / 16241 / 40 ] 2

for domain 1.shs.edu we select 190 heads / 1769 ( 10.74 )

for domain 1.shs.edu we kept 28 heads / 1769 ( 1.58 )

lemma POS F / dom F / o occ [ / dom / occ

éducation NC 0.013340 0.426 126 [ / 9445 / 296 ] 1

enseignement NC 0.012070 0.170 114 [ / 9445 / 671 ] 1

formation NC 0.011435 0.181 108 [ / 9445 / 596 ] 1

compétence NC 0.005823 0.185 55 [ / 9445 / 297 ] 3

école NC 0.004447 0.198 42 [ / 9445 / 212 ] 3

informatique NC 0.004341 0.506 41 [ / 9445 / 81 ] 1 old

didactique NC 0.004023 0.404 38 [ / 9445 / 94 ] 2

accompagnement NC 0.003812 0.303 36 [ / 9445 / 119 ] 1

situation NC 0.003388 0.153 32 [ / 9445 / 209 ] 2

enseignant NC 0.003282 0.419 31 [ / 9445 / 74 ] 1

pédagogie NC 0.003176 0.380 30 [ / 9445 / 79 ] 1

orientation NC 0.003070 0.201 29 [ / 9445 / 144 ] 1

connaissance NC 0.002541 0.157 24 [ / 9445 / 153 ] 1

autoformation NC 0.002329 0.880 22 [ / 9445 / 25 ] 1 old

scolarisation NC 0.002012 0.543 19 [ / 9445 / 35 ] 1 old

élève NC 0.001906 0.391 18 [ / 9445 / 46 ] 1

ordinateur NC 0.001800 0.447 17 [ / 9445 / 38 ] 1

geste NC 0.001694 0.158 16 [ / 9445 / 101 ] 1

partenariat NC 0.001588 0.179 15 [ / 9445 / 84 ] 1

professionnalisation NC 0.001588 0.221 15 [ / 9445 / 68 ] 1

professeur NC 0.001376 0.289 13 [ / 9445 / 45 ] 1

mathématique NC 0.001271 0.160 12 [ / 9445 / 75 ] 3

écrit NC 0.001165 0.149 11 [ / 9445 / 74 ] 1

obstacle NC 0.001165 0.208 11 [ / 9445 / 53 ] 1

tutorat NC 0.001059 0.625 10 [ / 9445 / 16 ] 1 old

étudiant NC 0.001059 0.149 10 [ / 9445 / 67 ] 1

stage NC 0.001059 0.333 10 [ / 9445 / 30 ] 1

e-inclusion NPP 0.001059 1.000 10 [ / 9445 / 10 ] 1 old

for domain 1.shs.litt we select 137 heads / 5127 ( 2.67 )

for domain 1.shs.litt we kept 52 heads / 5127 ( 1.01 )

lemma POS F / dom F / o occ [ / dom / occ

littérature NC 0.009945 0.594 142 [ / 14278 / 239 ] 1 old

roman NC 0.008475 0.742 121 [ / 14278 / 163 ] 1 old

écriture NC 0.006654 0.283 95 [ / 14278 / 336 ] 1

lecture NC 0.006303 0.156 90 [ / 14278 / 577 ] 1

théâtre NC 0.006163 0.371 88 [ / 14278 / 237 ] 2

figure NC 0.006093 0.172 87 [ / 14278 / 506 ] 1

poésie NC 0.005113 0.575 73 [ / 14278 / 127 ] 1 old

traduction NC 0.004903 0.255 70 [ / 14278 / 274 ] 2

voyage NC 0.004412 0.258 63 [ / 14278 / 244 ] 2

récit NC 0.004412 0.214 63 [ / 14278 / 295 ] 2

fiction NC 0.003852 0.404 55 [ / 14278 / 136 ] 1

poétique NC 0.003782 0.593 54 [ / 14278 / 91 ] 1 old

notice NC 0.003572 0.169 51 [ / 14278 / 302 ] 3

lettre NC 0.003152 0.285 45 [ / 14278 / 158 ] 2

réception NC 0.003082 0.177 44 [ / 14278 / 249 ] 1

voix NC 0.003012 0.274 43 [ / 14278 / 157 ] 2

réécriture NC 0.002872 0.612 41 [ / 14278 / 67 ] 1 old

portrait NC 0.002731 0.188 39 [ / 14278 / 208 ] 3

livre NC 0.002591 0.163 37 [ / 14278 / 227 ] 2

écrivain NC 0.002591 0.529 37 [ / 14278 / 70 ] 1 old

Proust NPP 0.002521 0.800 36 [ / 14278 / 45 ] 1 old

poète NC 0.002451 0.603 35 [ / 14278 / 58 ] 1 old

poème NC 0.002451 0.603 35 [ / 14278 / 58 ] 1 old

parole NC 0.002171 0.178 31 [ / 14278 / 174 ] 1

édition NC 0.002031 0.212 29 [ / 14278 / 137 ] 2

présence NC 0.001961 0.154 28 [ / 14278 / 182 ] 2

texte NC 0.001891 0.168 27 [ / 14278 / 161 ] 2

imaginaire NC 0.001821 0.164 26 [ / 14278 / 159 ] 1

correspondance NC 0.001821 0.250 26 [ / 14278 / 104 ] 1

personnage NC 0.001751 0.490 25 [ / 14278 / 51 ] 1

amour NC 0.001681 0.316 24 [ / 14278 / 76 ] 1

mort NC 0.001681 0.168 24 [ / 14278 / 143 ] 2

utopie NC 0.001611 0.160 23 [ / 14278 / 144 ] 1

métamorphose NC 0.001611 0.184 23 [ / 14278 / 125 ] 1

Montaigne NPP 0.001541 0.688 22 [ / 14278 / 32 ] 1 old

oeuvre NC 0.001541 0.151 22 [ / 14278 / 146 ] 2

Jean NPP 0.001541 0.179 22 [ / 14278 / 123 ] 2

voltaire NC 0.001471 0.700 21 [ / 14278 / 30 ] 1 old

manuscrit NC 0.001401 0.179 20 [ / 14278 / 112 ] 3

animal NC 0.001401 0.190 20 [ / 14278 / 105 ] 2

Alexandre NPP 0.001331 0.487 19 [ / 14278 / 39 ] 1

Flaubert NPP 0.001331 0.760 19 [ / 14278 / 25 ] 1 old

rire NC 0.001331 0.373 19 [ / 14278 / 51 ] 1

rêve NC 0.001261 0.234 18 [ / 14278 / 77 ] 1

métaphore NC 0.001261 0.200 18 [ / 14278 / 90 ] 2

Perceforest NPP 0.001191 1.000 17 [ / 14278 / 17 ] 1 old

auteur NC 0.001191 0.327 17 [ / 14278 / 52 ] 1

journal NC 0.001191 0.162 17 [ / 14278 / 105 ] 1

altérité NC 0.001121 0.340 16 [ / 14278 / 47 ] 1

réalisme NC 0.001121 0.364 16 [ / 14278 / 44 ] 2

avatar NC 0.001051 0.165 15 [ / 14278 / 91 ] 1

écho NC 0.001051 0.273 15 [ / 14278 / 55 ] 1

for domain 0.scco we select 183 heads / 1161 ( 15.76 )

for domain 0.scco we kept 9 heads / 1161 ( 0.78 )

lemma POS F / dom F / o occ [ / dom / occ

catégorisation NC 0.004776 0.203 15 [ / 3141 / 74 ] 1

psychologie NC 0.003184 0.169 10 [ / 3141 / 59 ] 2

stéréotype NC 0.002547 0.186 8 [ / 3141 / 43 ] 1

proverbe NC 0.002547 0.276 8 [ / 3141 / 29 ] 1

adjectif NC 0.002229 0.156 7 [ / 3141 / 45 ] 2

but NC 0.001910 0.333 6 [ / 3141 / 18 ] 1

ergonomie NC 0.001592 0.172 5 [ / 3141 / 29 ] 1

précocité NC 0.001592 1.000 5 [ / 3141 / 5 ] 1 old

distorsion NC 0.001273 0.333 4 [ / 3141 / 12 ] 2

for domain 1.shs.socio we select 182 heads / 5227 ( 3.48 )

for domain 1.shs.socio we kept 97 heads / 5227 ( 1.86 )

lemma POS F / dom F / o occ [ / dom / occ

exemple NC 0.016976 0.172 550 [ / 32398 / 3204 ] 1

enjeu NC 0.011760 0.181 381 [ / 32398 / 2110 ] 1

espace NC 0.007346 0.278 238 [ / 32398 / 855 ] 1

ville NC 0.007161 0.377 232 [ / 32398 / 615 ] 1

politique NC 0.007130 0.196 231 [ / 32398 / 1177 ] 1

construction NC 0.006636 0.165 215 [ / 32398 / 1303 ] 1

dynamique NC 0.006544 0.247 212 [ / 32398 / 857 ] 2

pratique NC 0.005988 0.181 194 [ / 32398 / 1074 ] 1

territoire NC 0.005988 0.412 194 [ / 32398 / 471 ] 1

géographie NC 0.005556 0.600 180 [ / 32398 / 300 ] 1 old

sociologie NC 0.005155 0.532 167 [ / 32398 / 314 ] 2 old

usage NC 0.005124 0.202 166 [ / 32398 / 823 ] 1

regard NC 0.005000 0.170 162 [ / 32398 / 951 ] 1

travail NC 0.004908 0.267 159 [ / 32398 / 595 ] 1

mobilité NC 0.004506 0.368 146 [ / 32398 / 397 ] 1

expérience NC 0.004445 0.159 144 [ / 32398 / 906 ] 1

place NC 0.004074 0.158 132 [ / 32398 / 833 ] 1

paysage NC 0.004043 0.322 131 [ / 32398 / 407 ] 1

forme NC 0.004013 0.177 130 [ / 32398 / 734 ] 1

réseau NC 0.003642 0.179 118 [ / 32398 / 661 ] 1

migration NC 0.003457 0.410 112 [ / 32398 / 273 ] 1

risque NC 0.003272 0.233 106 [ / 32398 / 455 ] 2

tourisme NC 0.002963 0.500 96 [ / 32398 / 192 ] 1 old

violence NC 0.002840 0.307 92 [ / 32398 / 300 ] 1

culture NC 0.002778 0.205 90 [ / 32398 / 439 ] 1

temps NC 0.002747 0.220 89 [ / 32398 / 404 ] 1

rapport NC 0.002716 0.175 88 [ / 32398 / 504 ] 2

enquête NC 0.002685 0.202 87 [ / 32398 / 431 ] 2

conflit NC 0.002685 0.262 87 [ / 32398 / 332 ] 2

trajectoire NC 0.002685 0.358 87 [ / 32398 / 243 ] 1

identité NC 0.002593 0.193 84 [ / 32398 / 435 ] 2

processus NC 0.002562 0.165 83 [ / 32398 / 502 ] 2

émergence NC 0.002562 0.180 83 [ / 32398 / 460 ] 2

transformation NC 0.002531 0.219 82 [ / 32398 / 374 ] 1

femme NC 0.002438 0.206 79 [ / 32398 / 384 ] 2

sport NC 0.002377 0.487 77 [ / 32398 / 158 ] 1

frontière NC 0.002377 0.297 77 [ / 32398 / 259 ] 1

changement NC 0.002284 0.173 74 [ / 32398 / 427 ] 1

acteur NC 0.002222 0.277 72 [ / 32398 / 260 ] 1

science NC 0.002191 0.166 71 [ / 32398 / 428 ] 2

parcours NC 0.002130 0.251 69 [ / 32398 / 275 ] 1

condition NC 0.002099 0.152 68 [ / 32398 / 447 ] 2

défi NC 0.002099 0.171 68 [ / 32398 / 397 ] 2

logique NC 0.002068 0.215 67 [ / 32398 / 312 ] 2

genre NC 0.002068 0.245 67 [ / 32398 / 273 ] 1

lieu NC 0.002068 0.235 67 [ / 32398 / 285 ] 2

corps NC 0.002037 0.163 66 [ / 32398 / 406 ] 1

jeu NC 0.002037 0.152 66 [ / 32398 / 433 ] 1

gouvernance NC 0.002006 0.185 65 [ / 32398 / 351 ] 2

mutation NC 0.002006 0.245 65 [ / 32398 / 265 ] 1

vulnérabilité NC 0.001975 0.457 64 [ / 32398 / 140 ] 1

cartographie NC 0.001945 0.268 63 [ / 32398 / 235 ] 1

dimension NC 0.001945 0.193 63 [ / 32398 / 327 ] 2

recomposition NC 0.001852 0.517 60 [ / 32398 / 116 ] 1 old

mobilisation NC 0.001852 0.324 60 [ / 32398 / 185 ] 2

entretien NC 0.001790 0.178 58 [ / 32398 / 325 ] 1

famille NC 0.001636 0.245 53 [ / 32398 / 216 ] 2

engagement NC 0.001605 0.259 52 [ / 32398 / 201 ] 1

patrimoine NC 0.001574 0.188 51 [ / 32398 / 271 ] 1

inégalité NC 0.001574 0.317 51 [ / 32398 / 161 ] 2

monde NC 0.001574 0.218 51 [ / 32398 / 234 ] 2

mode NC 0.001574 0.152 51 [ / 32398 / 336 ] 1

homme NC 0.001512 0.188 49 [ / 32398 / 260 ] 2

service NC 0.001482 0.224 48 [ / 32398 / 214 ] 1

agriculture NC 0.001482 0.222 48 [ / 32398 / 216 ] 3

paradoxe NC 0.001451 0.178 47 [ / 32398 / 264 ] 2

socialisation NC 0.001389 0.489 45 [ / 32398 / 92 ] 1

fabrique NC 0.001389 0.231 45 [ / 32398 / 195 ] 1

marché NC 0.001389 0.179 45 [ / 32398 / 251 ] 3

ressource NC 0.001358 0.181 44 [ / 32398 / 243 ] 1

entreprise NC 0.001327 0.167 43 [ / 32398 / 257 ] 3

mondialisation NC 0.001327 0.319 43 [ / 32398 / 135 ] 2

quartier NC 0.001327 0.426 43 [ / 32398 / 101 ] 3

Europe NPP 0.001296 0.203 42 [ / 32398 / 207 ] 2

accès NC 0.001296 0.164 42 [ / 32398 / 256 ] 2

fin NC 0.001296 0.156 42 [ / 32398 / 270 ] 2

France NPP 0.001266 0.192 41 [ / 32398 / 213 ] 3

transport NC 0.001235 0.231 40 [ / 32398 / 173 ] 2

école NC 0.001204 0.184 39 [ / 32398 / 212 ] 3

participation NC 0.001204 0.163 39 [ / 32398 / 239 ] 2

commerce NC 0.001204 0.235 39 [ / 32398 / 166 ] 2

transition NC 0.001204 0.148 39 [ / 32398 / 263 ] 2

invention NC 0.001142 0.167 37 [ / 32398 / 221 ] 2

voie NC 0.001142 0.160 37 [ / 32398 / 231 ] 1

lien NC 0.001142 0.154 37 [ / 32398 / 241 ] 1

aménagement NC 0.001080 0.230 35 [ / 32398 / 152 ] 1

eau NC 0.001080 0.241 35 [ / 32398 / 145 ] 1

prise NC 0.001080 0.170 35 [ / 32398 / 206 ] 2

savoir NC 0.001080 0.174 35 [ / 32398 / 201 ] 1

lutte NC 0.001080 0.175 35 [ / 32398 / 200 ] 2

population NC 0.001049 0.358 34 [ / 32398 / 95 ] 2

variabilité NC 0.001049 0.181 34 [ / 32398 / 188 ] 3

appropriation NC 0.001049 0.205 34 [ / 32398 / 166 ] 1

temporalité NC 0.001049 0.347 34 [ / 32398 / 98 ] 1

métier NC 0.001019 0.262 33 [ / 32398 / 126 ] 1

contexte NC 0.001019 0.151 33 [ / 32398 / 218 ] 1

association NC 0.001019 0.173 33 [ / 32398 / 191 ] 2

for domain 1.shs.geo we select 188 heads / 604 ( 31.13 )

for domain 1.shs.geo we kept 9 heads / 604 ( 1.49 )

lemma POS F / dom F / o occ [ / dom / occ

démographie NC 0.007557 0.200 9 [ / 1191 / 45 ] 1

appui NC 0.002519 0.250 3 [ / 1191 / 12 ] 1

fécondité NC 0.002519 0.200 3 [ / 1191 / 15 ] 1

excision NC 0.001679 0.667 2 [ / 1191 / 3 ] 1 old

végétation NC 0.001679 0.182 2 [ / 1191 / 11 ] 1

ADN NC 0.001679 0.333 2 [ / 1191 / 6 ] 1

vigie NC 0.001679 0.500 2 [ / 1191 / 4 ] 1 old

SIDA NPP 0.001679 0.667 2 [ / 1191 / 3 ] 1 old

mutilation NC 0.001679 0.500 2 [ / 1191 / 4 ] 1 old

for domain 1.shs.archeo we select 178 heads / 3430 ( 5.19 )

for domain 1.shs.archeo we kept 84 heads / 3430 ( 2.45 )

lemma POS F / dom F / o occ [ / dom / occ

site NC 0.013218 0.578 177 [ / 13391 / 306 ] 1 old

céramique NC 0.012546 0.828 168 [ / 13391 / 203 ] 1 old

archéologie NC 0.009559 0.547 128 [ / 13391 / 234 ] 2 old

donnée NC 0.008737 0.366 117 [ / 13391 / 320 ] 1

bilan NC 0.008065 0.161 108 [ / 13391 / 671 ] 1

résultat NC 0.007468 0.163 100 [ / 13391 / 613 ] 2

production NC 0.007244 0.179 97 [ / 13391 / 543 ] 2

fouille NC 0.006646 0.511 89 [ / 13391 / 174 ] 2 old

habitat NC 0.006497 0.420 87 [ / 13391 / 207 ] 1

occupation NC 0.006348 0.675 85 [ / 13391 / 126 ] 1 old

atelier NC 0.006049 0.409 81 [ / 13391 / 198 ] 1

dépôt NC 0.005675 0.576 76 [ / 13391 / 132 ] 1 old

décor NC 0.005153 0.575 69 [ / 13391 / 120 ] 2 old

sanctuaire NC 0.005003 0.779 67 [ / 13391 / 86 ] 1 old

nécropole NC 0.004406 0.787 59 [ / 13391 / 75 ] 1 old

industrie NC 0.004406 0.333 59 [ / 13391 / 177 ] 2

sépulture NC 0.004331 0.674 58 [ / 13391 / 86 ] 1 old

maison NC 0.003883 0.329 52 [ / 13391 / 158 ] 2

église NC 0.003809 0.293 51 [ / 13391 / 174 ] 3

découverte NC 0.003809 0.408 51 [ / 13391 / 125 ] 1

exploitation NC 0.003734 0.215 50 [ / 13391 / 233 ] 1

château NC 0.003435 0.374 46 [ / 13391 / 123 ] 2

monnaie NC 0.003435 0.359 46 [ / 13391 / 128 ] 3

ensemble NC 0.003360 0.413 45 [ / 13391 / 109 ] 1

campagne NC 0.003286 0.346 44 [ / 13391 / 127 ] 1

établissement NC 0.003136 0.400 42 [ / 13391 / 105 ] 1

inscription NC 0.003062 0.195 41 [ / 13391 / 210 ] 2

mobilier NC 0.002987 0.870 40 [ / 13391 / 46 ] 1 old

datation NC 0.002912 0.557 39 [ / 13391 / 70 ] 2 old

gisement NC 0.002838 0.543 38 [ / 13391 / 70 ] 2 old

peinture NC 0.002838 0.325 38 [ / 13391 / 117 ] 2

grotte NC 0.002688 0.507 36 [ / 13391 / 71 ] 1 old

prospection NC 0.002688 0.766 36 [ / 13391 / 47 ] 1 old

vestige NC 0.002464 0.635 33 [ / 13391 / 52 ] 1 old

témoignage NC 0.002390 0.198 32 [ / 13391 / 162 ] 2

faune NC 0.002390 0.582 32 [ / 13391 / 55 ] 2 old

reste NC 0.002315 0.620 31 [ / 13391 / 50 ] 1 old

tombe NC 0.002240 0.612 30 [ / 13391 / 49 ] 1 old

fortification NC 0.002166 0.509 29 [ / 13391 / 57 ] 2 old

monument NC 0.002166 0.333 29 [ / 13391 / 87 ] 1

peuplement NC 0.002016 0.474 27 [ / 13391 / 57 ] 1

âge NC 0.002016 0.250 27 [ / 13391 / 108 ] 1

vaisselle NC 0.002016 0.794 27 [ / 13391 / 34 ] 1 old

mission NC 0.001867 0.205 25 [ / 13391 / 122 ] 1

stèle NC 0.001867 0.610 25 [ / 13391 / 41 ] 1 old

fragment NC 0.001867 0.203 25 [ / 13391 / 123 ] 2

four NC 0.001792 0.686 24 [ / 13391 / 35 ] 1 old

artisanat NC 0.001718 0.657 23 [ / 13391 / 35 ] 1 old

chronologie NC 0.001718 0.426 23 [ / 13391 / 54 ] 1

vase NC 0.001643 0.733 22 [ / 13391 / 30 ] 1 old

enceinte NC 0.001643 0.550 22 [ / 13391 / 40 ] 1 old

terre NC 0.001643 0.265 22 [ / 13391 / 83 ] 1

mosaïque NC 0.001568 0.600 21 [ / 13391 / 35 ] 1 old

mine NC 0.001494 0.588 20 [ / 13391 / 34 ] 1 old

parure NC 0.001494 0.741 20 [ / 13391 / 27 ] 1 old

agglomération NC 0.001494 0.364 20 [ / 13391 / 55 ] 1

matière NC 0.001494 0.282 20 [ / 13391 / 71 ] 1

pierre NC 0.001419 0.268 19 [ / 13391 / 71 ] 1

indice NC 0.001419 0.153 19 [ / 13391 / 124 ] 1

temple NC 0.001419 0.442 19 [ / 13391 / 43 ] 1

thermes NC 0.001344 0.562 18 [ / 13391 / 32 ] 1 old

oppidum NC 0.001344 0.900 18 [ / 13391 / 20 ] 1 old

quartier NC 0.001344 0.178 18 [ / 13391 / 101 ] 3

tour NC 0.001344 0.188 18 [ / 13391 / 96 ] 1

monastère NC 0.001270 0.472 17 [ / 13391 / 36 ] 1

cathédrale NC 0.001270 0.395 17 [ / 13391 / 43 ] 2

néolithique NC 0.001195 0.800 16 [ / 13391 / 20 ] 1 old

inventaire NC 0.001195 0.150 16 [ / 13391 / 107 ] 1

animal NC 0.001195 0.152 16 [ / 13391 / 105 ] 2

témoin NC 0.001195 0.213 16 [ / 13391 / 75 ] 1

hache NC 0.001195 0.889 16 [ / 13391 / 18 ] 1 old

complexe NC 0.001195 0.281 16 [ / 13391 / 57 ] 1

abri NC 0.001195 0.800 16 [ / 13391 / 20 ] 1 old

chantier NC 0.001195 0.188 16 [ / 13391 / 85 ] 2

matériel NC 0.001120 0.577 15 [ / 13391 / 26 ] 1 old

amphore NC 0.001120 0.882 15 [ / 13391 / 17 ] 1 old

bâtiment NC 0.001120 0.385 15 [ / 13391 / 39 ] 1

villa NC 0.001120 0.484 15 [ / 13391 / 31 ] 1

meule NC 0.001120 0.789 15 [ / 13391 / 19 ] 1 old

fosse NC 0.001120 0.938 15 [ / 13391 / 16 ] 1 old

milieu NC 0.001045 0.156 14 [ / 13391 / 90 ] 1

bronze NC 0.001045 0.636 14 [ / 13391 / 22 ] 1 old

enclos NC 0.001045 0.667 14 [ / 13391 / 21 ] 1 old

faciès NC 0.001045 0.700 14 [ / 13391 / 20 ] 1 old

for domain 0.chim we select 173 heads / 785 ( 22.04 )

for domain 0.chim we kept 44 heads / 785 ( 5.61 )

lemma POS F / dom F / o occ [ / dom / occ

synthèse NC 0.045387 0.218 123 [ / 2710 / 565 ] 2

matériau NC 0.015867 0.231 43 [ / 2710 / 186 ] 3

catalyseur NC 0.014022 0.826 38 [ / 2710 / 46 ] 1 old

catalyse NC 0.013284 0.923 36 [ / 2710 / 39 ] 1 old

préparation NC 0.011070 0.283 30 [ / 2710 / 106 ] 2

oxydation NC 0.010701 0.617 29 [ / 2710 / 47 ] 1 old

chimie NC 0.008487 0.365 23 [ / 2710 / 63 ] 1

spectroscopie NC 0.006642 0.173 18 [ / 2710 / 104 ] 2

nanoparticule NC 0.006273 0.472 17 [ / 2710 / 36 ] 1

polymère NC 0.005904 0.533 16 [ / 2710 / 30 ] 1 old

réactivité NC 0.004428 0.414 12 [ / 2710 / 29 ] 1

membrane NC 0.004059 0.500 11 [ / 2710 / 22 ] 1 old

ligand NC 0.004059 0.786 11 [ / 2710 / 14 ] 1 old

hydrogénation NC 0.003690 1.000 10 [ / 2710 / 10 ] 1 old

activation NC 0.003321 0.237 9 [ / 2710 / 38 ] 1

polymérisation NC 0.002952 0.615 8 [ / 2710 / 13 ] 1 old

molécule NC 0.002583 0.219 7 [ / 2710 / 32 ] 1

oxyde NC 0.002583 0.700 7 [ / 2710 / 10 ] 1 old

nanocomposite NC 0.002214 0.375 6 [ / 2710 / 16 ] 1

rhéologie NC 0.001845 0.200 5 [ / 2710 / 25 ] 1

spéciation NC 0.001845 0.625 5 [ / 2710 / 8 ] 1 old

greffage NC 0.001845 0.556 5 [ / 2710 / 9 ] 1 old

alliage NC 0.001845 0.500 5 [ / 2710 / 10 ] 1 old

filtration NC 0.001845 0.312 5 [ / 2710 / 16 ] 2

chromatographie NC 0.001476 0.500 4 [ / 2710 / 8 ] 1 old

?RMN NC 0.001476 0.571 4 [ / 2710 / 7 ] 1 old

adsorption NC 0.001476 0.190 4 [ / 2710 / 21 ] 1

carbène NC 0.001476 1.000 4 [ / 2710 / 4 ] 1 old

acidité NC 0.001476 1.000 4 [ / 2710 / 4 ] 1 old

hydroconversion NC 0.001476 1.000 4 [ / 2710 / 4 ] 1 old

?Chimisorption NPP 0.001476 1.000 4 [ / 2710 / 4 ] 1 old

composé NC 0.001107 0.250 3 [ / 2710 / 12 ] 1

fluor NC 0.001107 0.600 3 [ / 2710 / 5 ] 1 old

résine NC 0.001107 0.375 3 [ / 2710 / 8 ] 1

click NC 0.001107 1.000 3 [ / 2710 / 3 ] 1 old

réacteur NC 0.001107 0.176 3 [ / 2710 / 17 ] 1

photochimie NC 0.001107 0.500 3 [ / 2710 / 6 ] 1 old

confinement NC 0.001107 0.167 3 [ / 2710 / 18 ] 1

enthalpie NC 0.001107 0.600 3 [ / 2710 / 5 ] 1 old

immobilisation NC 0.001107 0.429 3 [ / 2710 / 7 ] 1

dimère NC 0.001107 0.600 3 [ / 2710 / 5 ] 1 old

hydroformulation NC 0.001107 1.000 3 [ / 2710 / 3 ] 1 old

hydrodésulfuration NC 0.001107 0.750 3 [ / 2710 / 4 ] 1 old

Peg NPP 0.001107 1.000 3 [ / 2710 / 3 ] 1 old

for domain 0.sdu we select 195 heads / 1240 ( 15.73 )

for domain 0.sdu we kept 35 heads / 1240 ( 2.82 )

lemma POS F / dom F / o occ [ / dom / occ

variabilité NC 0.007619 0.149 28 [ / 3675 / 188 ] 3

géologie NC 0.004626 0.586 17 [ / 3675 / 29 ] 1 old

enregistrement NC 0.003810 0.233 14 [ / 3675 / 60 ] 1

gisement NC 0.003537 0.186 13 [ / 3675 / 70 ] 2

datation NC 0.003265 0.171 12 [ / 3675 / 70 ] 2

bassin NC 0.003265 0.293 12 [ / 3675 / 41 ] 1

faune NC 0.003265 0.218 12 [ / 3675 / 55 ] 2

légende NC 0.003265 0.185 12 [ / 3675 / 65 ] 1

ammonite NC 0.002993 1.000 11 [ / 3675 / 11 ] 1 old

excursion NC 0.002993 0.647 11 [ / 3675 / 17 ] 1 old

érosion NC 0.002721 0.323 10 [ / 3675 / 31 ] 1

volcanisme NC 0.002177 0.615 8 [ / 3675 / 13 ] 1 old

sédimentation NC 0.001905 0.318 7 [ / 3675 / 22 ] 1

massif NC 0.001905 0.700 7 [ / 3675 / 10 ] 1 old

karst NC 0.001633 0.240 6 [ / 3675 / 25 ] 2

calcaire NC 0.001361 0.833 5 [ / 3675 / 6 ] 1 old

littoral NC 0.001361 0.152 5 [ / 3675 / 33 ] 1

sédiment NC 0.001361 0.333 5 [ / 3675 / 15 ] 1

feuille NC 0.001361 0.357 5 [ / 3675 / 14 ] 1

forage NC 0.001361 1.000 5 [ / 3675 / 5 ] 1 old

stratigraphie NC 0.001361 0.312 5 [ / 3675 / 16 ] 1

ambre NC 0.001361 0.556 5 [ / 3675 / 9 ] 1 old

métamorphisme NC 0.001361 1.000 5 [ / 3675 / 5 ] 1 old

volcan NC 0.001361 0.312 5 [ / 3675 / 16 ] 1

géothermie NC 0.001361 0.833 5 [ / 3675 / 6 ] 1 old

réchauffement NC 0.001088 0.200 4 [ / 3675 / 20 ] 1

terrasse NC 0.001088 0.267 4 [ / 3675 / 15 ] 1

brachiopode NC 0.001088 1.000 4 [ / 3675 / 4 ] 1 old

vertébré NC 0.001088 0.800 4 [ / 3675 / 5 ] 1 old

granite NC 0.001088 0.667 4 [ / 3675 / 6 ] 1 old

désagrégation NC 0.001088 0.444 4 [ / 3675 / 9 ] 1

sondage NC 0.001088 0.167 4 [ / 3675 / 24 ] 1

précipitation NC 0.001088 0.267 4 [ / 3675 / 15 ] 1

géomorphologie NC 0.001088 0.200 4 [ / 3675 / 20 ] 1

monitoring NC 0.001088 0.250 4 [ / 3675 / 16 ] 1

for domain 1.shs.art we select 164 heads / 3366 ( 4.87 )

for domain 1.shs.art we kept 44 heads / 3366 ( 1.31 )

lemma POS F / dom F / o occ [ / dom / occ

art NC 0.012781 0.197 111 [ / 8685 / 563 ] 2

musique NC 0.011054 0.366 96 [ / 8685 / 262 ] 2

vitrail NC 0.009211 0.941 80 [ / 8685 / 85 ] 1 old

notice NC 0.007830 0.225 68 [ / 8685 / 302 ] 3

théâtre NC 0.006678 0.245 58 [ / 8685 / 237 ] 2

portrait NC 0.006102 0.255 53 [ / 8685 / 208 ] 3

peinture NC 0.005987 0.444 52 [ / 8685 / 117 ] 2

cinéma NC 0.003915 0.366 34 [ / 8685 / 93 ] 1

collection NC 0.003685 0.262 32 [ / 8685 / 122 ] 1

décor NC 0.003569 0.258 31 [ / 8685 / 120 ] 2

artiste NC 0.003454 0.492 30 [ / 8685 / 61 ] 1

sculpture NC 0.003339 0.569 29 [ / 8685 / 51 ] 1 old

oeuvre NC 0.003224 0.192 28 [ / 8685 / 146 ] 2

église NC 0.002994 0.149 26 [ / 8685 / 174 ] 3

jardin NC 0.002533 0.272 22 [ / 8685 / 81 ] 1

musée NC 0.002418 0.221 21 [ / 8685 / 95 ] 1

dessin NC 0.002303 0.385 20 [ / 8685 / 52 ] 1

tableau NC 0.002073 0.247 18 [ / 8685 / 73 ] 1

photographie NC 0.002073 0.191 18 [ / 8685 / 94 ] 1

iconographie NC 0.002073 0.277 18 [ / 8685 / 65 ] 1

manuscrit NC 0.002073 0.161 18 [ / 8685 / 112 ] 3

verrière NC 0.001957 1.000 17 [ / 8685 / 17 ] 1 old

peintre NC 0.001957 0.630 17 [ / 8685 / 27 ] 1 old

scène NC 0.001842 0.211 16 [ / 8685 / 76 ] 1

opéra NC 0.001727 0.500 15 [ / 8685 / 30 ] 1 old

hôtel NC 0.001727 0.375 15 [ / 8685 / 40 ] 1

chantier NC 0.001727 0.176 15 [ / 8685 / 85 ] 2

restauration NC 0.001612 0.167 14 [ / 8685 / 84 ] 2

film NC 0.001382 0.158 12 [ / 8685 / 76 ] 1

répertoire NC 0.001382 0.261 12 [ / 8685 / 46 ] 1

cathédrale NC 0.001267 0.256 11 [ / 8685 / 43 ] 2

maître NC 0.001267 0.244 11 [ / 8685 / 45 ] 1

architecte NC 0.001267 0.324 11 [ / 8685 / 34 ] 2

danse NC 0.001267 0.224 11 [ / 8685 / 49 ] 1

dramaturgie NC 0.001267 0.440 11 [ / 8685 / 25 ] 1

miroir NC 0.001151 0.182 10 [ / 8685 / 55 ] 1

recueil NC 0.001151 0.182 10 [ / 8685 / 55 ] 1

renaissance NC 0.001036 0.164 9 [ / 8685 / 55 ] 1

fortune NC 0.001036 0.176 9 [ / 8685 / 51 ] 1

estampe NC 0.001036 0.900 9 [ / 8685 / 10 ] 1 old

musicien NC 0.001036 0.265 9 [ / 8685 / 34 ] 1

compositeur NC 0.001036 0.818 9 [ / 8685 / 11 ] 1 old

salon NC 0.001036 0.643 9 [ / 8685 / 14 ] 1 old

concert NC 0.001036 0.600 9 [ / 8685 / 15 ] 1 old

for domain 1.shs.psy we select 242 heads / 939 ( 25.77 )

for domain 1.shs.psy we kept 29 heads / 939 ( 3.09 )

lemma POS F / dom F / o occ [ / dom / occ

psychologie NC 0.007510 0.339 20 [ / 2663 / 59 ] 2

psychanalyse NC 0.004506 0.300 12 [ / 2663 / 40 ] 1

autisme NC 0.004131 0.407 11 [ / 2663 / 27 ] 1

clinique NC 0.003380 0.250 9 [ / 2663 / 36 ] 1

croyance NC 0.003004 0.154 8 [ / 2663 / 52 ] 1

addiction NC 0.002629 0.318 7 [ / 2663 / 22 ] 1

psychologue NC 0.002253 0.750 6 [ / 2663 / 8 ] 1 old

adolescence NC 0.002253 0.353 6 [ / 2663 / 17 ] 1

hallucination NC 0.001878 0.500 5 [ / 2663 / 10 ] 1 old

psychose NC 0.001878 0.417 5 [ / 2663 / 12 ] 1

acceptation NC 0.001878 0.185 5 [ / 2663 / 27 ] 1

fratrie NC 0.001878 0.385 5 [ / 2663 / 13 ] 1

Freud NPP 0.001502 0.308 4 [ / 2663 / 13 ] 1

subjectivation NC 0.001502 0.364 4 [ / 2663 / 11 ] 1

symptôme NC 0.001502 0.211 4 [ / 2663 / 19 ] 1

douleur NC 0.001502 0.222 4 [ / 2663 / 18 ] 1

attachement NC 0.001502 0.190 4 [ / 2663 / 21 ] 1

sevrage NC 0.001502 0.571 4 [ / 2663 / 7 ] 1 old

scarification NC 0.001127 0.750 3 [ / 2663 / 4 ] 1 old

estime NC 0.001127 0.750 3 [ / 2663 / 4 ] 1 old

cognition NC 0.001127 0.250 3 [ / 2663 / 12 ] 1

jouissance NC 0.001127 0.273 3 [ / 2663 / 11 ] 1

dangerosité NC 0.001127 0.375 3 [ / 2663 / 8 ] 1

psychopathologie NC 0.001127 0.429 3 [ / 2663 / 7 ] 1

psychothérapie NC 0.001127 0.429 3 [ / 2663 / 7 ] 1

psychodrame NC 0.001127 1.000 3 [ / 2663 / 3 ] 1 old

distorsion NC 0.001127 0.250 3 [ / 2663 / 12 ] 2

questionnaire NC 0.001127 0.167 3 [ / 2663 / 18 ] 1

trauma NC 0.001127 0.429 3 [ / 2663 / 7 ] 1

for domain 1.shs.scipo we select 212 heads / 2507 ( 8.46 )

for domain 1.shs.scipo we kept 32 heads / 2507 ( 1.28 )

lemma POS F / dom F / o occ [ / dom / occ

élection NC 0.008110 0.533 80 [ / 9864 / 150 ] 1 old

sociologie NC 0.005474 0.172 54 [ / 9864 / 314 ] 2

démocratie NC 0.005474 0.284 54 [ / 9864 / 190 ] 2

Europe NPP 0.005069 0.242 50 [ / 9864 / 207 ] 2

mobilisation NC 0.004765 0.254 47 [ / 9864 / 185 ] 2

parti NC 0.004663 0.535 46 [ / 9864 / 86 ] 1 old

justice NC 0.004663 0.176 46 [ / 9864 / 262 ] 3

État NC 0.003954 0.231 39 [ / 9864 / 169 ] 2

institutionnalisation NC 0.002332 0.271 23 [ / 9864 / 85 ] 1

décentralisation NC 0.002129 0.344 21 [ / 9864 / 61 ] 1

région NC 0.002028 0.215 20 [ / 9864 / 93 ] 1

gouvernement NC 0.002028 0.312 20 [ / 9864 / 64 ] 1

politisation NC 0.002028 0.435 20 [ / 9864 / 46 ] 1

citoyenneté NC 0.001926 0.229 19 [ / 9864 / 83 ] 1

Turquie NPP 0.001825 0.450 18 [ / 9864 / 40 ] 1

vote NC 0.001622 0.356 16 [ / 9864 / 45 ] 1

parlement NC 0.001622 0.254 16 [ / 9864 / 63 ] 2

maire NC 0.001521 0.600 15 [ / 9864 / 25 ] 1 old

islam NC 0.001419 0.292 14 [ / 9864 / 48 ] 1

Afrique NPP 0.001318 0.203 13 [ / 9864 / 64 ] 1

syndicat NC 0.001318 0.265 13 [ / 9864 / 49 ] 1

européanisation NC 0.001217 0.286 12 [ / 9864 / 42 ] 1

opinion NC 0.001217 0.245 12 [ / 9864 / 49 ] 1

immigration NC 0.001217 0.158 12 [ / 9864 / 76 ] 1

gauche NC 0.001217 0.545 12 [ / 9864 / 22 ] 1 old

droite NC 0.001217 0.480 12 [ / 9864 / 25 ] 1

Tunisie NPP 0.001115 0.478 11 [ / 9864 / 23 ] 1

métropole NC 0.001115 0.200 11 [ / 9864 / 55 ] 2

syndicalisme NC 0.001014 0.333 10 [ / 9864 / 30 ] 1

député NC 0.001014 0.588 10 [ / 9864 / 17 ] 1 old

Amérique NPP 0.001014 0.196 10 [ / 9864 / 51 ] 1

nationalisme NC 0.001014 0.312 10 [ / 9864 / 32 ] 1

## A.2.2 Liste des têtes transdisciplinaires

Le tableau suivant présente nos 93 têtes transdisciplinaires. Est indiqué le lemme, la catégorie du discours, si le lemme appartient aux formes du lexique transdisciplinaire des écrits scientifiques (Tutin, 2008) (LTES) et si le lemme appartient à la liste des signalling nouns (Flowerdew et Forest, 2015) (SIGN) avec la fréquence normalisée dans leur corpus. Nous notons que :

* 21 des têtes transdisciplinaires n’appartiennent pas au LTES, 72 y appartiennent, soit 77 %.
* 11 des têtes transdisciplinaires n’appartiennent pas aux signalling nouns, 61 y appartiennent, soit 88 %.

**N° LEMME POS LTES SIGN**

1 étude NC LTES 18

2 analyse NC LTES 178

3 approche NC LTES 246

4 cas NC LTES 890

5 exemple NC LTES 421

6 enjeu NC

7 modèle NC LTES 474

8 évolution NC LTES 39

9 rôle NC LTES 153

10 apport NC LTES 16

11 réflexion NC LTES 16

12 évaluation NC LTES 10

13 représentation NC LTES 11

14 effet NC LTES 393

15 application NC LTES 44

16 construction NC LTES 2

17 introduction NC LTES 70

18 question NC LTES 313

19 histoire NC 20

20 outil NC LTES 7

21 développement NC LTES 39

22 état NC LTES 10

23 impact NC 96

24 modélisation NC LTES 3

25 système NC LTES 109

26 regard NC 5

27 contribution NC LTES 16

28 pratique NC LTES 73

29 recherche NC LTES 2

30 comparaison NC LTES 44

31 relation NC LTES 93

32 mise NC

33 dynamique NC

34 méthode NC LTES 280

35 influence NC LTES 44

36 espace NC LTES

37 place NC LTES 21

38 stratégie NC LTES 205

39 perspective NC LTES 36

40 expérience NC LTES 3

41 forme NC LTES 88

42 projet NC LTES 37

43 élément NC LTES 33

44 utilisation NC LTES 5

45 retour NC 29

46 aspect NC LTES 78

47 essai NC 41

48 usage NC LTES 73

49 politique NC 2

50 enseignement NC

51 réseau NC LTES 7

52 présentation NC LTES 11

53 interaction NC LTES 15

54 gestion NC LTES

55 lecture NC LTES 33

56 note NC 13

57 processus NC LTES 230

58 formation NC

59 dispositif NC LTES 7

60 principe NC LTES 251

61 science NC

62 proposition NC LTES 46

63 remarque NC 21

64 théorie NC LTES 494

65 bilan NC LTES 83

66 structure NC LTES 13

67 mesure NC LTES 46

68 problème NC LTES 619

69 émergence NC

70 point NC 393

71 enquête NC LTES 16

72 intérêt NC LTES 29

73 intégration NC LTES

74 organisation NC LTES 13

75 conception NC 28

76 résultat NC LTES 572

77 travail NC LTES 24

78 image NC LTES 5

79 concept NC LTES 143

80 défi NC 26

81 variation NC LTES 15

82 production NC LTES 2

83 méthodologie NC 13

84 cadre NC LTES 31

85 activité NC LTES 59

86 changement NC LTES 209

87 figure NC LTES 88

88 culture NC

89 discours NC 51

90 objet NC LTES 3

91 perception NC LTES 85

92 traitement NC LTES 300

93 rapport NC LTES 10

# A3. Étiquettes utilisées par Talismane et HAL

## A3.1 Catégories morphosyntaxiques de Talismane

Ces informations sont tirées de <http://joliciel-informatique.github.io/talismane/#tagset>.

|  |  |
| --- | --- |
| **Code** | **Catégorie morphosyntaxique** |
| ADJ | Adjectif |
| ADV | Adverbe |
| ADVWH | Adverbe interrogatif |
| CC | Conjonction de coordination |
| CLO | Clitique objet |
| CLR | Clitique réflexif |
| CLS | Clitique sujet |
| CS | Conjonction de subordination |
| DET | Déterminant |
| DETWH | Déterminant interrogatif |
| ET | Mot étranger |
| I | Interjection |
| NC (que nous rassemblons dans NOUN) | Nom commun |
| NPP (que nous rassemblons dans NOUN) | Nom propre |
| P (que nous rassemblons dans PREP) | Préposition |
| P+D (que nous rassemblons dans PREP) | Préposition et déterminant combinés (“du”) |
| P+PRO (que nous rassemblons dans PREP) | Préposition et pronom combiné (“duquel”) |
| PONCT | Ponctuation |
| PRO | Pronom |
| PROREL | Pronom relatif |
| PROWH | Pronom interrogatif |
| V (que nous rassemblons dans VERB) | Verbe à l’indicatif |
| VIMP (que nous rassemblons dans VERB) | Verbe à l’impératif |
| VINF (que nous rassemblons dans VERB) | Verbe à l’infinitif |
| VPP (que nous rassemblons dans VERB) | Verbe au participe paéssé |
| VPR (que nous rassemblons dans VERB) | Verbe au participe présent |
| VS (que nous rassemblons dans VERB) | Verbe au subjonctif |

## A3.2 Relations de dépendance de Talismane

TODO

A3.3 Code des 27 disciplines de HAL retenues

|  |  |  |
| --- | --- | --- |
| 01 | 0.chim | Chimie |
| 02 | 0.info | Informatique |
| 03 | 0.math | Mathématiques |
| 04 | 0.phys | Physique |
| 05 | 0.qfin | Économie et finance quantitative |
| 06 | 0.scco | Sciences cognitives |
| 07 | 0.sde | Sciences de l'environnement |
| 08 | 0.sdu | Planète et Univers |
| 09 | 0.sdv | Sciences du Vivant |
| 10 | 1.shs.anthro | Anthropologie |
| 11 | 1.shs.archeo | Archéologie et Préhistoire |
| 12 | 1.shs.archi | Architecture |
| 13 | 1.shs.art | Art et histoire de l'art |
| 14 | 1.shs.autre | Autres |
| 15 | 1.shs.droit | Droit |
| 16 | 1.shs.edu | Éducation |
| 17 | 1.shs.geo | Géographie |
| 18 | 1.shs.gestion | Gestion et management |
| 19 | 1.shs.hist | Histoire |
| 20 | 1.shs.infocom | Sciences de l'information et de la communication |
| 21 | 1.shs.ling | Linguistique |
| 22 | 1.shs.litt | Littératures |
| 23 | 1.shs.phil | Philosophie |
| 24 | 1.shs.psy | Psychologie |
| 25 | 1.shs.scipo | Science politique |
| 26 | 1.shs.socio | Sociologie |
| 27 | NONE | Pas de discipline associée |

# A4. Éléments techniques

## A4.A Présentation de l’API de requêtage de notre corpus

Nous présentons dans cette partie notre interface de programmation de l’application (API) que nous avons dévelopée afin d’intérroger notre corpus.

Requêtes sur notre corpus pour filtrer le corpus, trouver des titres et faire des statistiques.

stat('domain')

Produit un comptage des titres selon la discipline des titres. Le résultat est un dictionnaire où la clé est la discipline et la valeur le nombre de titre dans cette discipline.

stat(('nb\_parts', 'nb\_segments'))

Produit un comptage des titres selon les combinaisons des valeurs possibles pour le nombre de parties et le nombre de segments. Le résultat est un dictionnaire où la clé est un tuple constitué d’une combinaison existante de valeurs des deux dimensions, par exemple 1 partie, 2 segments, et la valeur le nombre de titre correspondant à cette combinaison, le nombre de titres ayant 1 partie et 2 segments.

count({'nb\_parts' : 1, 'nb\_segments' : 2})

Compte le nombre de titre ayant une partie et deux segments.

t12 = select({'nb\_parts' : 1, 'nb\_segments' : 2})

Création d’un sous-corpus composé des titres ayant une partie et deux segments. On peut ensuite utiliser les requêtes stat et count sur celui-ci via une variable globale qui contient le corpus courant.

find({'nb\_roots' : 2}, nb=20)

Cherche et affiche 20 titres ayant 2 racines.

find({'roots.0.lemma' : 'rôle', 'roots.1.lemma' : 'cas',   
 'segments.0.lemma' : '.'})

Cherche et affiche 5 titres dont la tête du premier segment est le lemme *rôle*, celle du second segment le lemme *cas* et dont le signe de ponctuation segmentant est un point. Cette requête ne marche que sur un corpus constitué de titres à au moins deux segments.

avg('nb\_segments')

minn('nb\_segments')

maxx('nb\_segments')

Obtient respectivement la moyenne des valeurs, la valeur minimum et la valeur maximum pour la clé *nb\_segments* dans le corpus actuel.

## A4.B Analyse de 100 titres traités par Talismane

Nous avons analysé 100 titres traités par Talismane pour vérifier qu’il catégorisait bien les têtes de segments. Nous prenons 20 titres de chaque structure (nombre de segments et position des racines dans les segments) qui nous intéresse. Nous indiquons :

* Son identifiant dont la couleur indique le résultat de l’analyse pour le titre :
  + en **vert** si le titre a été analysé correctement en ce qui concerne la détection de têtes de segments,
  + en **orange** si l’analyse est discutable mais n’impacte pas notre analyse,
  + en **rouge** si elle est fausse en ne détecant pas la bonne tête de segment,
  + en **violet** si la promotion d’un mot en tête de segment par notre algorithme fait changer le titre de catégorie structuelle,
  + en **rose** si une tête n’a pas été détectée.
* Pour les cinq structures qui nous intéressent, un code segment-racine de la forme :
  + 1\_\_ pour un titre ayant 1 segment et 1 racine,
  + 2\_\_ pour un titre ayant 1 segment et 2 racines,
  + 1:0 pour un titre ayant 1 racine dans son premier segment et 0 dans son second,
  + 0:1 pour l’inverse,
  + 1:1 pour un titre ayant 1 racine dans chacun de ses deux segments.
* Les têtes de segment sont en gras et :
  + en **vert** si elles sont correctement catégorisées et lemmatisées,
  + en **bleu** si le lemme est incorrect ou inconnu (lemme ignoré pour NPP),
  + en **orange** si la catégorie est incorrecte,
  + en **rouge** s’il ne s’agit pas d’une racine,
  + en **violet** si elles ne sont pas détectées par Talismane mais par notre algorithme,
  + en **rose** si elles ne sont pas détectées ni par Talismane ni par notre algorithme.

-------------------------------------------------------------------------------------------------

001 **62230** 1\_\_ Un possible **modele** semiotique global de la communication

*Note 01 : L’absence d’accent fait que Talismane n’associe pas ce NC au lemme modèle.*

002 **62250** 1\_\_ L'**IMPACT** DE L'EDITION ELECTRONIQUE SUR LA CRISE DU KOSOVO

003 **460613** 1\_\_ Un **indicateur** de politique d'ouverture à l'immigration

004 **62244** 1\_\_ Le **déplacement** médiatique du débat politique

005 **110369** 1\_\_ L'**imprimerie** et sa diffusion en Extrême-Orient

006 **911256** 1\_\_ Les **enfants** d'Hygie

007 **410464** 1\_\_ **Optimisation** de la précipitation des métaux lourds en mélange

008 **911470** 1\_\_ L'**héritage** du Boiteux d'Orgemont

009 **216325** 1\_\_ **DIFFUSION** INTERGRANULAIRE ET ÉNERGIE DES JOINTS DE GRAINS

010 **760276** 1\_\_ **Dépôt** sec des aérosols à l'interface air-eau

011 **1808328** 1\_\_ **Modélisation** de la structure d'un mélange à haute dilution

012 **1015139** 1\_\_ **Analyse** écophysiologique de la nitrophilie des espèces adventices

013 **264210** 1\_\_ Un **regard** sur les approches basées sur la vision par ordinateur

014 **1759146** 1\_\_ L’**implantation** de l’abbaye de Conques dans les environs de Sainte-Foy-la-Grande

au XIe siècle

015 **215986** 1\_\_ La **persistance** du droit successoral de l'Ancien Régime dans l'Europe du XIXe

siècle

*Note 02 : On remarque que Talismane fait dépendre le* du *de* persistance *plutôt que* Europe *mais*

*cela n’affecte pas notre analyse qui se limite à la tête de segment.*

016 **162355** 1\_\_ **Faut**-il jeter la Méditerranée avec l'eau du bain ?

017 **215983** 1\_\_ La **défense** de la victime en France au XIXe et au XXe siècle

018 **110374** 1\_\_ **Rédaction** de 120 notices

019 **62249** 1\_\_ **Vers** une approche ethnographique des usages des Technologies de l'Information et

de la Communication au sein des petites et moyennes entreprises malaisiennes

*Note 03 : L’enchaînement de compléments de nom peut perde Talismane : il ne sait plus par quoi*

*est régi la préposition* de*. Ici celui avant l’Information est indiqué comme étant*

*régi par* approche *au lieu de* Technologies*. Cela n’a pas d’incidence sur notre*

*travail.*

020 **1808326** 1\_\_ **Algorithme** de construction de modèles markoviens multidimensionnels pour le

mélange des poudres

-------------------------------------------------------------------------------------------------

021 **216380** 2\_\_ **DIFFUSION** AVANT ET **ARRIÈRE** D'IONS LOURDS ET MOMENTS ANGULAIRES COMPLEXES

022 **1258669** 2\_\_ **Contenu** et **exigences** du travail

*Note 04 : Talismane normalement ne désigne que le premier NC d’un schéma NC CC NC comme*

*tête. Ici, il désigne les deux NC ce qui n’est pas cohérent.*

023 **312877** 2\_\_ **Demain** la **géographie** sociale.

*Note 05 : La promotion de l’adverbe comme racine est discutable.*

024 **1015192** 2\_\_ **Évaluation** de la dispersion des propriétés mécaniques d'un matériau composite par

sous-**échantillonnage**

*Note 06 : La présence d’un tiret provoque une erreur dans Talismane.*

025 **1808361** 2\_\_ Conditionnement des boues par gel-**dégel**

*Note 07 :* dégel *est désigné comme racine alors que ce n’est clairement pas le cas à cause du*

*tiret.*

026 **264579** 2\_\_ **Institutions** [Les **humanités** et les grandes institutions du savoir en France]

*Note 08 : On peut considérer le texte entre crochets comme un segment non détecté.*

027 **1258688** 2\_\_ **Comparaison** isoenzymatique de deux populations boliviennes (altitude et plaine)

**de** Triatoma infestans (Hemiptera\, Reduviidae)

*Note 09 :* de *est désigné comme racine alors que ce n’est clairement pas le cas.*

028 **162715** 2\_\_ **Transfert** de chaleur et de masse dans une salle d'opérations conditionnée\,

**comparaison** entre deux modes de soufflage

*Note 10 : La virgule n’est pas considérée comme segmentante mais ici elle devrait l’être.*

029 **264613** 2\_\_ **Accès** à l'information et reconnaissance d'un droit à l'information

environnementale - Le nouveau **contexte** juridique international

*Note 11 : Le tiret n’est pas considéré comme segmentant mais ici il devrait l’être. Cela*

*est facilité par la présence d’une majuscule.*

030 **62420** 2\_\_ **De** l'appropriation inachevée du concept de genre (gender) **en** communication

organisationnelle

*Note 12 :* en *est désigné comme racine alors que ce n’est clairement pas le cas.*

031 **216445** 2\_\_ **APPLICATION** DES MÉTHODES STATISTIQUES AU CALCUL DES CHAMPS THERMIQUES TURBULENTS

NON **HOMOGÈNES**

*Note 13 :* HOMOGÈNES *est désigné comme racine alors que ce n’est clairement pas le cas.*

032 **960687** 2\_\_ **Amitiés**\, des **sciences** sociales aux réseaux sociaux de l'internet

033 **216532** 2\_\_ **TRANSITION** MÉTAL-**SEMICONDUCTEUR** DANS LES COMPOSÉS Cr2S3-xSex ET Cr2+εSe3

*Note 14 : La présence d’un tiret provoque une erreur dans Talismane.*

034 **1609898** 2\_\_ Les **Vigiles** **debout**

*Note 15 : Talismane ne devrait prendre que le verbe conjugué.*

035 **960764** 2\_\_ **Misère** de l'hyper-**spécialisation** et dérives du professionnalisme

*Note 16 : La présence d’un tiret provoque une erreur dans Talismane.*

036 **62668** 2\_\_ **Bibliothèques** numériques et Google-**Print**

*Note 17 :* Print *est désigné comme racine alors que ce n’est clairement pas le cas.*

037 **1559698** 2\_\_ Dispositif **de de** caractérisatioon simultanée de l'abondance de pucerons et de la

croissance végétative d'arbres fruitiers

*Note 18 : La répétition de la préposition* de *entraîne une erreur dans Talismane.*

038 **264587** 2\_\_ Le **jeu**\, une **approche** philosophique

*Note 19 : ici, la virgule a une valeur segmentante.*

039 **460685** 2\_\_ Surveillance de chorégraphies de Web Services basées sur WS-**CDL**

*Note 20 : La présence d’un tiret provoque une erreur dans Talismane.*

040 **62434** 2\_\_ **Développement** stratégique du tourisme sportif de rivière par régulation

corporatiste L'**expérience** du bassin de Saint Anne (Québec) appliquée aux Rivières

de Provence

*Note 21 : Oubli d’un point entre les deux segments du titre. La présence d’une majuscule*

*permet de bien repéréer la segmentation manquante.*

-------------------------------------------------------------------------------------------------

041 **62397** 1:0 **Réinterroger** les structures documentaires : **de** la numérisation à

l'informatisation

042 **62226** 1:0 Les **temporalités** médiatiques des personnes âgées : des **évolutions** dans la

stabilité

043 **360068** 1:0 La **performativité** de l'évidence : **analyse** du discours néolibéral

*Note 22 : Le mot n’est pas rattaché à son lemme par Talismane car son statut lexical est*

*discutable.*

044 **1061179** 1:0 La **Société** de la Carte géologique de France (1869-1872) : une éphémère **réaction** à

la création du Service de la Carte géologique de la France

045 **360074** 1:0 **Dynamique** technologique controversée et débat démocratique : le **cas** des micros et

nanotechnologies

046 **62256** 1:0 **Traces** de contenus africains sur Internet : **entre** homogénéité et identité

047 **216312** 1:0 **MODÈLES** THÉOTIQUES DE LA STRUCTURE DES JOINTS DE GRAINS.LES **MODÈLES** DE STRUCTURE

DES JOINTS DE GRAINS ET LEUR UTILISATION

*Note 23 : Les deux têtes sont les mêmes.*

048 **1759477** 1:0 Les **objets** communicants\, La **problématique** des Antennes: **Dispositif** **pour** détecter

le vêlage des vaches.

*Note 24 :* pour *est détecter faussement par notre algorithme comme un mot à promouvoir en racine*

*car* Dispositif *et* pour *sont régis par* objets*. De plus, on a une virgule segmentante,*

*la majuscule qui la suit montrant clairement le début d’un segment. Il s’agit donc*

*d’un titre à trois segments.*

049 **760329** 1:0 L'**omniprésence** de la famille au sein de l'exploitation agricole : une **situation**

de fait encouragé par les règles de droit

050 **1208785** 1:0 **SymbAphidBase** : une **base** de données nouvelle dédiée aux symbiotes de pucerons

pour stocker et visualiser les génomes séquencés en standardisant leurs

annotations

051 **264568** 1:0 **Bill** Viola : **voir** l'eau ou la transparence en mouvement

*Note 25 :* Bill *est caractérisé comme un NC au lieu d’un NPP.*

052 **1759420** 1:0 Les **objets** communicants\, La **problématique** des Antennes; **Balises** de Détresse

*Note 26 : trois problèmes dans ce titre :* problématique *est considérée comme un adjectif, la*

*virgule n’est pas segmentante mais ici elle l’est, et Balises est détecté par notre*

*algorithme. En fait, il s’agit un titre à trois segments et non deux.*

053 **460618** 1:0 **PERCEPTION** DE L'INDÉPENDANCE DE L'AUDITEUR : **ANALYSE** PAR LA THÉORIE D'ATTRIBUTION

054 **1707597** 1:0 **Élites** maléfiques et ""complot pédophile"" : **paniques** morales autour des enfants

055 **1759142** 1:0 **Formation** et évolution des paroisses de la basse vallée du Drot : **essai** de

synthèse

056 **859899** 1:0 **Classification** floue généralisée : **Application** à la quantification de la stéatose

sur des images histologiques couleurs

057 **510693** 1:0 Les **gastroentérites** aiguës à rotavirus de l'enfant : une **priorité** de santé

publique.

058 **960530** 1:0 **Monde** pluriel : **penser** l'unité des sciences sociales

059 **659177** 1:0 **Reconnaissance** et appropriation : **pour** une anthropologie du travail

060 **62190** 1:0 **Métiers** émergents de la nouvelle économie: **identification** des compétences

attendues et typologie des métiers exercés

-------------------------------------------------------------------------------------------------

061 **1660207** 0:1 Quel **pouvoir** de stabilisation à l’échelle de l’UEM : le pacte de stabilité et de

croissance **est**-il viable ?

062 **659285** 0:1 L'**Etat** et les "" autres "" : **comparer** la visibilisation de la main-d'œuvre

immigrée

063 **62609** 0:1 Le **Libre** Accès (Open Access) : **partager** les résultats de la recherche

*Note 27 :* Libre *est caractérisé comme NPP ainsi que* Accès*. On peut se poser la question si*

*ce n’est pas le syntagme nominale entier* Libre Accès *qui devrait être racine.*

064 **960680** 0:1 **De** l'apprenti footballeur **au** petit-rat de l'Opéra : comment les institutions

d'excellence **agissent** face aux dispositions sociales des apprentis ?

*Note 28 : Notre algorithme devrait se contenter de ne prendre que* de.

065 **1258715** 0:1 **Référentiels** de compétences : ce que l'instrument **fait** à la logique compétence

066 **860275** 0:1 La **question** périurbaine : la **repenser** en tenant enfin compte de ce qui motive les

périurbains

067 **62568** 0:1 **Transférabilité** des connaissances : une re-conceptualisation de la distinction

tacite / **explicite**

*Note 29 : Talismane catégorise* explicite *comme V au lieu d’ADJ. De ce fait, il désigne*

explicite *comme tête au lieu* de re-conceptualisation.

068 **264762** 0:1 **Théophile** Gautier : **Regardez**\, mais ne touchez pas (comédie)

*Note 30 : On peut se poser la question si ce n’est pas le syntagme nominal entier*

Théophile Gautier *qui devrait être pris comme tête par notre algorithme.*

069 **1015049** 0:1 Les (**il**)**légalités** ambiguës dans le travail policier : comment l'espace **devient**

prétexte

*Note 31 : l’utilisation du suffixe entre parenthèses* il *perd Talisman. Il le catégorise comme*

*CLS. Notre algorithme ensuite trouve deux mots à prendre pour têtes au lieu d’un.*

070 **1358243** 0:1 **Evolution** de l'arboricolie chez les Cercopithèques: analyse **combinée** de données

moléculaires\, morpho-anatomiques et comportementales

*Note 32 :* combinée *est choisi comme racine alors qu’*analyse *devrait l’être.*

071 **1061109** 0:1 **ImPAC** Lyon : **évaluer** l'impact environnemental et thermique de l'exploitation des

aquifères superficiels pour la climatisation

072 **1759247** 0:1 **Relation** image/**son** : de l'illustration **sonore** à la fusion multi-modale

*Note 33* : sonore *est caractérisé comme V au lieu de ADJ et comme tête alors que de*

*est* de *est un meilleur candidat. On remarque la construction* de X à Y.

*Notre algorithme propose* Relation *est bien la tête du premier segment et*

*incorrectement* son *qui est mal catégorisé : DET au lieu de NC.*

073 **760065** 0:1 **D'**une catastrophe\, l'**autre** : **vivre** avec l'atome

*Note 34 : Notre algorithme détecterait* autre *également comme tête car il est régi par vivre.*

*Mais nous limitons notre algorithme à ne prendre que le premier mot comme racine.*

074 **110247** 0:1 **Vers** une économie des fonctionnalités: **changer** nos rapports avec le produit pour

des économies d'échelle et des nouvelles logiques de responsabilités

075 **809358** 0:1 **Après** la délocalisation...les PME **doivent**-elles relocaliser ?

07 6 **460346** 0:1 Une jeune **fille** changée en jeune homme : homélie sur un miracle survenu dans le

monastère **couvent** de Qartmin\, dans le Tur Abdin

*Note 35 : Erreur classique de confondre le NC* couvent *avec le V couvrir, de plus il ne s’agit*

*pas de la tête de segment,* homélie *y prêtant plus sûrement.*

078 **1060698** 0:1 **Extension** de procédure: ""Le législateur nous **garde** de l'opportunité du juge

079 **312714** 0:1 **Mise** au point sur ""Les cathares devant l'histoire"" et retour sur ""L'histoire

du catharisme en discussion: le débat sur la charte de Niquinta n'est pas **clos**

*Note 36 :* Mise*, détecté par notre algorithme, est catégorisé comme VPP au lieu de NC.*

080 **162674** 0:1 **Communication** financière : quelles **sont** les pratiques des entreprises ?

-------------------------------------------------------------------------------------------------

081 **1258625** 1:1 Un nouvel **OVNI** dans le ciel réunionnais : la **transparence** des prix

082 **62241** 1:1 **De** l'anarchisme au combat identitaire : l'**internet** comme média révolutionnaire ?

083 **62366** 1:1 **Communication** et changement organisationnel : le **concept** de chaîne   
 d'appropriation

084 **264580** 1:1 Mystique et **magie** naturelle : les **paysages** mystiques de l'Espagne

*Note 37 :* Mystique *est catégorisée comme ADJ, Talismane privilégie donc le NC* magie *comme*

*racine. Mais il aurait dû soit choisir* Mystique.

085 **216338** 1:1 **MIGRATION** DES JOINTS DE GRAINS.LA **MIGRATION** DES JOINTS INTERGRANULAIRES

*Note 38 : La capitalisation ne pose pas de problème à Talismane. Les deux têtes sont le même*

*mot.*

086 **1609872** 1:1 La **création** d'entreprise en réponse au rêve d'île : l'**ambivalence** d'une   
 attractivité fondée sur le cadre de vie.

087 **659340** 1:1 **Mise** à disposition des données géologiques de surface : **Création** d'un accès sous   
 InfoTerre

*Note 39 : la nominalisation de la locution verbale “mettre à disposition” n’est pas bien   
 catégorisée.*

088 **960668** 1:1 **Brevet** et patrimoine génétique : la **brevetabilité** des organismes génétiquement

modifiés

089 **62616** 1:1 **Projet** DigiCulture : **pour** un portrait des usages et des usagers des ressources

culturelles numériques canadiennes

090 **62386** 1:1 **PRATIQUES** ENONCIATIVES HYPERTEXTUELLES : **VERS** DE NOUVELLES ORGANISATIONS

MEMORIELLES.

091 **110466** 1:1 L**'avenir** de la Common law en français : un **point** de vue d'Europe continentale

092 **1109003** 1:1 **Estimation** des quantiles conditionnels par quantification optimale : nouveaux

**résultats**

093 **1108914** 1:1 **Présentation** d'une langue: le **hongrois**

094 **609991** 1:1 **Variation** du risque de cancer du sein en fonction de la nature de la mutation du

gène ATM. **Étude** familiale rétrospective

095 **62386** 1:1 **PRATIQUES** ENONCIATIVES HYPERTEXTUELLES : **VERS** DE NOUVELLES ORGANISATIONS

MEMORIELLES.

096 **1015246** 1:1 L'**impact** des enceintes urbaines médiévales sur le territoire et ses limites.

L'**exemple** de la Lorraine et de l'Alsace

097 **1258763** 1:1 **Phèdre** janséniste ? **retour** sur un lieu commun (2)

*Note 40 :* Phèdre *n’est pas catégorisé comme un NPP mais comme un NC.*

098 **1409780** 1:1 **Développement** et politique. Le **cas** d'une politique de santé en Géorgie.

099 **62382** 1:1 Quels **modèles** pour la publication sur le web? Le **cas** des contenus informationnels

et culturels.

*Note 41 : Talismane arrive à scinder le* ? *du mot* web.

100 **560355** 1:1 Un **tournant** participatif ? Une **mise** en perspective historique de la participation

du public dans les politiques scientifiques américaines

*Note 42 : Ici, mise est bien reconnu comme une nature nominale.*

# A5. Index des tableaux

TODO

# A6. Index des graphiques

TODO

# A7. Index des notions mentionnées

TODO